

## Generic Structure Storage and Retrieval

MICHAEL F. LYNCH,\* JOHN M. BARNARD, and STEPHEN M. WELFORD

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

Received March 7, 1985

The role of generic structures in the chemical knowledge base is described, with particular reference to patents. Operational information systems providing access to generic structures are reviewed, and past and present research leading to improved services is described.

### INTRODUCTION

Since the middle of the 19th century, when the empirical chemical arts began to give rise to useful new products, patent protection of chemical inventions has been sought and granted. One of the earliest instances was the synthetic dyestuff mauve obtained by Perkin<sup>1</sup> through the chance reaction of aniline with sulfuric acid and potassium dichromate, for which he applied for protection in a provisional specification first submitted in 1856. The full British Patent No. 1984 of 1856, sealed on February 20, 1857, includes not only aniline, but also toluidine, xylidene, and cumidine, so that this patent must also be considered as constituting the verbal expression of a generic chemical structure, although its publication predated Kekule's formulation of the structure of benzene by a decade.<sup>2</sup>

It is in the context of chemical patents that generic chemical structures find their strongest forms of development, although similar forms of expression are used widely by chemists in both written and spoken communications, as evidenced by expressions such as alkyltoluidines, acyl halides, and the like.

Chemical patents, frequently including generic chemical structures, provide a rich source of information, in terms both of new science and of new technology. Their principal purpose is to afford protection to inventors, for some period of time, for commercial exploitation of the fruits of their innovation and investment. Patents are granted predominantly on the authority vested in a patent office by individual states, in return for the publication of the new knowledge gained and of means of expressing this knowledge.

Patents thus comprise two aspects: first, the description of the invention and of procedures that a person skilled in the art may apply to effect them, and second, the delineation of the rights of monopoly in exploitation held by the patentee, as expressed in the claims. In the first respect, patents are often the first, and sometimes the only, form of publication of novel chemical science. In the second respect, they are legal documents, the interpretation of which is often a matter for resolution by courts of law.

Since patents are issued predominantly by individual states (an exception being European patents), which differ in what they accept as patentable, in whether or not they examine the submissions for validity or novelty, and in terms of the time that elapses between submission and publication, the oversight of the patent literature is a highly complex matter, in which secondary information services play a dominant role, since bibliographic control on an international basis is necessary. However, the matter is simplified somewhat for the information services as the Patent Convention of 1883 allows patents to be grouped into "families", all members of which claim the same "priority date". Secondary information services generally identify as the "basic" patent the patent for the country (or "major" country) in which it is first published and all other appearances of it as "equivalents".<sup>3,4</sup>

The numbers of chemical patents issued annually has roughly kept pace over the years with the increase in the

number of other types of chemical publications, although changes in practice, both in terms of national patent offices and the patent laws under which they operate, and also of the coverage of secondary information services make exact counting difficult.

### SIGNIFICANCE OF CHEMICAL PATENTS

It is the chemical industry and in particular its pharmaceutical and agricultural chemical branches, but also, increasingly, biotechnology and protein engineering, that perceive the crucial role of patent information in their activities. It is well established that academic chemists often fail to appreciate fully the fact that much novel chemistry appears only in patents, since a relatively small proportion of patents results in corresponding publications in the nonpatent literature,<sup>5</sup> and that literature searches which ignore patent sources are likely to be incomplete.

It has been estimated that in the pharmaceutical industry a novel substance synthesized for activity screening has a 1 in 10 000 chance of being successfully marketed, while the average investment, over an 8-12-year period, for each marketed product has been assessed at between \$50 and \$200 million.<sup>6</sup>

The purposes for which searches based on molecular structure are performed are diverse.<sup>7</sup> In the first instance, the requirement for novelty in an invention calls for a review of all existing published sources, in the widest sense, for any information that may be seen to anticipate the invention and that could thus lead to the failure of an application or its invalidation through subsequent litigation. Searches for this purpose are carried out as soon as a promising "lead" compound has been identified and before resources are invested in the detailed examination of the substance and its analogues.

The results of such searches may lead to validity searches, where the purpose is to set aside an existing patent owned by a competitor, by demonstrating that prior art rendered the obstructive patent invalid. An associated purpose is the infringement search, which is undertaken to determine whether competitors' products encroach on a company's patents, or vice versa.

### CHARACTERISTICS OF GENERIC CHEMICAL STRUCTURES

Generic structures are also known as Markush structures, after the inventor who in 1925 successfully challenged the U.S. Patent Office's rejection of his application for a patent for a class of novel pyrazolone dyes, opening the way for wide acceptance of this genus.<sup>8</sup>

Generic structures cover a wide spectrum of forms of expression. The central feature of a generic structure is the notion of an invariant part or parts, usually expressed as a structure diagram (although this may occasionally be vestigial), and associated variable groups usually comprising lists of



Michael Lynch was awarded B.Sc. and Ph.D. degrees in chemistry by the National University of Ireland, after studying at University College Dublin. As an undergraduate he held a studentship at the Northrhine-Westphalian Institute of Technology at Aachen; he subsequently held a postdoctoral fellowship, awarded by the Royal Commission for the Exhibition of 1851, at the Swiss Federal Institute of Technology, where he worked with Professor V. Prelog. After 2 years in research in industry in Britain, he joined Chemical Abstracts Service, and later took charge of the Basic Research Department. He has spent the past 20 years at Sheffield University's Department of Information Studies, where his research interests include the identification of interesting structural manipulations of text and chemical structure representations.



John Barnard graduated from Birmingham University in 1976 with a B.Sc. in biochemistry and subsequently obtained an M.Sc. in information studies from Sheffield University. After working for 1 year in the pharmaceutical industry, he returned to the Department of Information Studies at Sheffield, where he has spent 6 years as a member of the generic chemical structures project team, being awarded a Ph.D. in 1983. He has also been active in the Chemical Structure Association [formerly the Chemical Notation Association (U.K.)], serving on its Executive Committee in 1980–1984 and as Chairman in 1983–1984.



Stephen Welford graduated from the University of Durham in 1977 with a B.Sc. in chemistry and from the University of Sheffield in 1979 with an M.Sc. in information studies. He has since worked in Sheffield University's Department of Information Studies, researching into the computer processing of generic chemical structures for information retrieval and into applications of artificial intelligence techniques in chemical information science. He was awarded a Ph.D. in 1983.

alternatives that themselves are expressed as partial structure diagrams, line formulas, radicals named with specific or generic nomenclature, or verbal descriptions based on properties or uses. The associated radicals may be attached at fixed or variable locations on the invariant part. Groups may occur with varying incidences, either within given ranges or without any specification of numbers. Variables may be further substituted by defined or by undefined groups. Again, groups may be independent or may combine, for instance, to form an additional ring. The selection of combinations may be arbitrary or may be subject to express logical conditions, which, for example, prohibit certain combinations of structural variables where these constitute prior art.

All of these features contrast strongly with specific chemical structures, in which the description of constituent atoms and bonds is fully determinate, at least to the level of the two-dimensional topology.

A frequent circumstance is the citation of numeric or verbal expressions qualifying generic radical names and delineating, for instance, permissible numbers of atoms of carbon or other elements, features such as the type and number of unsaturations, ring numbers and their sizes, and whether or not ring systems may include heteroatoms or delocalized bonds. Instances of these include the following: 3–5C alkyl group; *n*-alkyl group; unsubstituted 5–10C cycloalkyl group. Generic structures thus encompass families of structures, the extent of which may be small in number or large, but bounded (in that all instances may be exhaustively generated from that expression) or unbounded in that one or more of the radical expressions is unbounded in its possible extensions.<sup>9,10</sup> A related problem, that of incompletely specified substances, has recently been systematically studied by Gordon.<sup>11,12</sup>

This natural complexity of denotation is compounded further by the fact that patent specifications frequently include claims at differing levels of specificity. Thus, the first claim is likely to use the most generic terminology. Subsequent claims may instance the radicals to be understood as defined for the purpose of that patent by those terms, while one or more particular substances will be claimed specifically, some or all of which have been prepared and characterized. Thus, the first claim may include the definition of a variable group by the expression "aryl"; further claims may limit this to include only phenyl or naphthyl groups, and the characterized or exemplified substances may have phenyl and 1-naphthyl substituents at the point in question. This characteristic of generic structures in patents has recently been discussed by Hyams.<sup>13</sup>

All of these factors combine to make the matter of storage and representation of chemical structural aspects of generic structures a problem of some magnitude. Solutions leading to operational information services have thus far been based predominantly on the approximation provided by the application of a variety of fragmentation codes.

#### EXISTING GENERIC STRUCTURE INFORMATION SYSTEMS

Patent offices generally publish patents in an official gazette, using procedures developed for purposes of standardized description through the Patent Cooperation Treaty, the World International Property Organisation (WIPO), and the Committee for International Cooperation in Information Retrieval among Patent Offices (ICIREPAT). Examining Patent Offices, in particular, maintain large organized collections of patent and literature sources. These serve the patent examiners in their determination of the acceptability of new applications and employ a variety of conventional classification and indexing techniques as well as computer-based retrieval systems.

The U.S. Patent and Trademark Office has recently embarked on a major process of modernization, with the intention

of automating essential activities by the end of the century.<sup>14,15</sup> This project envisages the replacement of paper-based files by electronic data bases and communications techniques to support its task of processing over 100 000 patent and 60 000 trademark applications each year. The progress of this enterprise will clearly be followed with great interest by patent users as well as by other patent offices around the world.

The patent and literature collections, and associated retrieval systems of national patent offices, while invaluable to those who can readily visit and consult them, are subordinate in importance to the commercially based secondary information services specializing in, or providing extensive coverage of, chemical patent information. These services have been extensively discussed in the literature.<sup>16-20</sup>

Foremost among the information services providing retrieval capabilities in respect of generic chemical structures is Derwent Publications Ltd. Its principal service is the *World Patents Index*. This includes the *Central Patents Index* (CPI), offering complete coverage of chemical technology inventions from 29 patent-issuing authorities and comprising a wide range of alerting and retrospective search services in printed, microform, and machine-readable forms.<sup>21,22</sup> Three sections within CPI, FARMDOC, AGDOC, and CHEMDOC, cover pharmaceutical, agrochemical, and general chemical patents, respectively. In addition to a wide variety of subject index and classification terms, generic chemical structures in patents in these sections are described in terms of the CPI code.

This fragmentation code originated in 1963 for use in the FARMDOC service, the first of these sections. It was based originally on the 80-column punched card; superimposed codes representing fragments occurring within the structure description were punched into the card bearing the printed Derwent abstract, so that searches for fragments or combinations of fragments produced the card and the abstract, incorporating the generic structure description. The code was later extended for use with AGDOC and CHEMDOC, and the means of storage and search were upgraded to include first magnetic tape searching and, subsequently, on-line searching with conventional inverted file software. Further additional features include *Ring Index* numbers to characterize the ring systems included within generic chemical structures and some thousands of specific substances whose mention in a patent is specifically indicated.

The CPI code, which has undergone many revisions during its life, is a manually assigned code. Each fragment number (no longer limited to the 960 punch positions of a punch card) corresponds to a functional group, ring system, or other feature of chemical significance, including generic features such as "spiro-ring fused heterocycle" and "aromatic thioether". Features are generally coded in a superimposed manner, whether present in the invariant part of a generic structure or in variables. While careful encoding and checking ensures high levels of recall in Boolean searches, the relevance level is often low.<sup>23-27</sup>

Today, searches of CPI are available from System Development Corp. (SDC), Dialog Information Services, and the Paris-based Télésystèmes-Questel. A recently developed microcomputer-based Chemical Code Menu Program is available to assist users in deciding which codes to use for particular structural concepts.

In contrast, the GREMAS code devised by Fugmann<sup>28,29</sup> constitutes an open-ended fragment code whose terms are constructed and assigned by rule. It is the basis for the services provided, predominantly to the West German chemical industry, by Internationale Dokumentationsgesellschaft für Chemie m.b.H. (IDC). Since the three-letter codes that form the basis of structural description are synthetic in their derivation, more than 10 000 of these are possible. They are, furthermore, complemented by a wider range still of syntactic

codes, and variable and invariant parts of the generic structure are themselves differentiated, so that both in description and in search substantial degrees of differentiation can be achieved, although at significant costs in terms of the level of expertise of the coders and search intermediaries. Under an agreement with Derwent Publications Ltd., IDC uses Derwent material as the basis for the creation of its search files.

At substantially lower levels of description, both the American Petroleum Institute, through its APIPAT system,<sup>30</sup> and the IFI-Plenum Data Co., through its CLAIMS file,<sup>31</sup> index generic chemical structures in patents. In that the substances of interest in APIPAT—often monomers—are less complex than those of interest to the pharmaceutical industry, the variety of terms is small and, as regards functionality, emphasizes unsaturation and heteroatom types. On the other hand, the IFI/Plenum fragmentation code, used to index U.S. patents from 1950 to date, has a wider range of fragment types. The most detailed structure indexing is available in the Comprehensive Data Base, where the fragment code, comprising over 10 000 terms, is used to describe generic structures and specific substances not themselves used as index terms. In the case of generic structures, the fragments are differentiated according to whether they occur in invariant ("must" terms) or variable ("possible" terms) parts of the structure, and the search algorithms reflect this difference.<sup>32</sup>

## RESEARCH ON IMPROVED METHODS OF STORING GENERIC STRUCTURES FOR SEARCH

As the new generation of specific chemical substance information systems exemplified by CAS Online, COUSIN, MACCS, and DARC<sup>33</sup> is welcomed by users, so the demand for graphic structure access to generic chemical structure files has grown. Indeed, the ability to pose queries in generic structure form is already a feature of certain of these systems. In 1979, Silk<sup>34</sup> outlined possible developments of existing systems that could lead to improved structure searching of the patent literature.

Ray and Kirsch,<sup>35</sup> in 1957, first described a method for storing a topological description of specific chemical substances in a form in which an atom-by-atom search could determine the presence or absence of substructures within the molecule. Soon after this, Opler's work on the display of organic structures on a cathode ray tube appeared.<sup>36</sup> These early studies prompted a spate of developments in the early 1960s, from which the specific chemical structure information systems we use today have been derived.

The earliest work considering the problem of computer handling of generic structures dates from 1958 and largely followed the tradition of applying fragmentation codes, which was to be the norm for practice for at least the next quarter of a century. Leibowitz, Frome, and Andrews<sup>37</sup> described initial collaborative work between the U.S. Patent and Trademark Office and the National Bureau of Standards on the Variable Scope Search System (VS3); this was based on a hierarchical fragmentation of specific chemical structures into ring and chain components and identification of functional groups, all represented by means of an open-ended fragmentation code, with punched-card methods for storage and search. Related work by Frome examined systems of less general scope, in particular, for steroids and for phosphorus compounds, which were used on a trial basis within the Patent Office.<sup>38</sup>

Quite different in nature, and based on Ray and Kirsch's work, the HAYSTAC system<sup>39,40</sup> had as its objective the storage of topological representations of generic structures. The requirements for search, including the need for search screens, were explicitly discussed, but the work, though imaginative, was premature, not least because of the scale and

speed of the hardware available at that time. SEAC (Standard's Electronic Automatic Computer) had 2048 44-bit words of storage (1536 words of high-speed memory, 1024 words of electrostatic, and 512 words of mercury-delay line memory) while its tape drives with fixed tapes had a density of only 200 bits per in. Even in 1967, formats for representing generic structures were still under discussion by this group,<sup>41</sup> although the machines then available included the NBS PILOT, as well as IBM 1410 and 360/40 machines.

Meanwhile, at the West German chemical company BASF AG, Meyer had begun experiments on the input and storage of organic structures, including certain types of generic structures.<sup>42</sup> By 1966, the work had reached a stage where the Formula Reading Machine and substructure search of files indexed by the GREMAS code (and a further code for dyestuffs) were in routine use. The Formula Reading Machine is still in use 20 years later (although it has been replaced at BASF by sophisticated keyboard input) for encoding both specific structures and generic structures, the latter including up to nine variables at each of three different attachment points. Sophisticated search routines are available for interactive search at the fragment search level, or in batch mode for atom-by-atom searches, on substantial files of generic structures.<sup>43</sup>

Linear notations have also been examined as a method of representation of generic structures. Thus, Silk devised and used a notation system to describe characteristic groups including generic radicals, resulting from the fragmentation of generic structures; an in-house system was operated on this basis for agrochemical research.<sup>44</sup> Sneed et al.<sup>45</sup> examined extension of the Hayward notation to deal with determinate generic structures. Similar work was pursued on the Wiswesser Line Notation by Fraser-Williams (Scientific Systems) Ltd., as reported by Jackson,<sup>46</sup> while Krishnamurthy and Lynch examined ALWIN from this viewpoint.<sup>47</sup>

#### UNIVERSITY OF SHEFFIELD PROJECT

The joint work between Krishnamurthy and Lynch was the point of departure for the Sheffield project, which has been in progress since 1979;<sup>48-55</sup> its importance lay in the introduction of formal linguistic theory as a powerful tool to assist in the solution of the complex representational problems.

The purpose of the project is to examine the requirements for future generic structure storage and search systems and to address the theoretical and practical issues needing solution before such systems, based primarily on a complete topological record of the generic structure rather than exclusively on degenerate, fragment-based representations, can be put into operation.

From the outset it was clear that a powerful new toolkit of representational methods and search algorithms, and computing resources well beyond those in use in chemical information systems today, would be required to complement the knowledge and experience gained from the operation of specific chemical structure information systems and to provide practicable and cost-effective solutions.

At an early stage in this work, it became evident that the notation-based approach first studied was inherently inflexible and ill suited to the task of mirroring accurately the varied components of generic structures. For instance, the requirement for a contiguous series of letter locants for the peripheral atoms of a ring system would call for separate labelings for ring systems with one or more rings of variable size, or, indeed, where options exist for the formation of additional rings.

The first requirement is for a flexible mechanism to mirror as closely as possible all the features encountered in generic structures as outlined earlier. The mechanism must support data-base creation operations, as well as query formulation,

by providing an easy mapping between the forms of expression of generic structures in patents and query formulations and their representation in machine-readable form.<sup>48</sup> The same criterion is important from the user's viewpoint, in that it should be readily understandable. Thus, it must be hospitable to partial structure diagrams for the invariant part and for variables in this form, as well as to line-formulas and specific and generic radical names, together with the numeric or other qualifiers used with generic nomenclature. It should also support the expression of logical relationships identified within the patent specification. It must serve as the medium for storage of an archival data base, from which other functions, in particular the task of searching, can readily be supported.

The mechanism developed for this purpose is GENSAL (GENeric Structure LAnguage).<sup>49</sup> It has been shown to be easy to learn and to apply to a wide variety of patents; its coverage, in terms of the proportion of generic structures in typical patents to which it has been applied successfully, is extremely high, and there are few circumstances to which it is not adaptable. Over 2000 generic structures from recent chemical patents have been encoded in GENSAL; these records now form a test data base for the development and evaluation of search algorithms. Figure 1 shows the Abstract of a Patent Specification prepared by Derwent Publications Ltd., and Figure 2 shows the GENSAL representation of the generic structure it describes.

Formally, GENSAL has a context-free grammar, which may, like that of modern programming languages, be shown by means of a set of syntax diagrams or set of Backus-Naur Form (BNF) production rules. The program that processes it, called the GENSAL Interpreter Program,<sup>53</sup> operates in much the same way as a programming language compiler, checking the syntactic and semantic correctness of the GENSAL statements and generating an internal "machine-level" representation of the generic structure.

This latter is the Extended Connection Table Representation (ECTR) and is a complex network of partial connection tables and other records, linked together to show the logical, positional, and multiplicative relationships between them.<sup>51</sup> It is intended to be transparent to the user, but like GENSAL, it is a complete and unambiguous representation of the generic structure.

The GENSAL record for the expression of generic structures is checked interactively at the time of input for ostensible correctness on the basis of criteria such as observance of the defined syntax and of normal valencies. Alternatively, the record can be checked in batch mode, and editing corrections can be made subsequently in the light of error messages. This record is the archival form, not least because of its relative compactness, due to use of specific and generic radical names and line formulas. Consistency checking involves the creation of the Extended Connection Table Representation, in which all radicals represented as names or line formulas for convenience in input are expanded to partial connection tables.

Generic radical terms, which cannot be stored as partial connection tables, are stored in the ECTR as a list of structural parameters, with appropriate values or ranges of values, which together provide an intensional description of the class of radicals. The structural parameters presently in use are listed in Table I, while additional parameters that have been recommended for inclusion in the system are listed in Table II. The input system recognizes default values for these parameters for a number of common generic radical terms, while provision is made for the selective replacement of these values by the analyst during input of the GENSAL record.

The ECTR is relatively large, is time consuming to create, and uses dynamic storage, which means that no arbitrary limits are set on the size or complexity of generic structures that can be handled. It is not seen, for the moment, as the primary

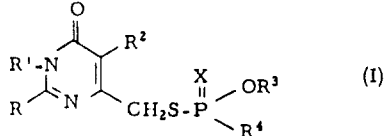
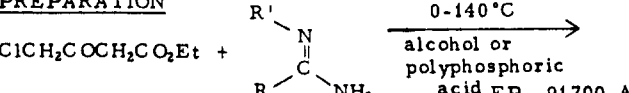
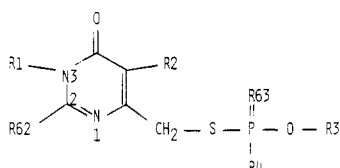
<p>83-796654/43 C01 USRU 10.12.81          UNIROYAL INC *EP --91-700-A          00.00.83-EP-106372 (+US-329157) (19.10.83) A01n-57/24 C07f-09/65          Pesticidal s-pyrimidinyl:methyl phosphoro-thioate ester(s) - useful as insecticides, nematocides and acaricides</p>	<p>C(5-B1M, 12-B2, 12-B4, 12-N1, 12-N2) 1013          R<sup>3</sup> is 1-3C alkyl; and          R<sup>4</sup> is 1-3C alkoxy or propylthio; provided that if R is phenyl, R<sup>3</sup> is Me and R<sup>4</sup> is MeO).</p>
<p>C83-102753 D/S: AT BE CH DE FR GB IT LI LU NL SE          S-Pyrimidinylmethyl phosphorothioates of formula (I) and their hydrochlorides are new</p> <div style="text-align: center;">  <p>(I)</p> </div> <p>(R is 1-3C alkyl or phenyl;          R' is H; or          R and R' together form (a) 1,3-butadiene-1,4-diyl opt. substd. by Cl or Me, or (b) 1,2-ethenediylthio where S is attached to C;          R<sup>2</sup> is H or halogen;          X is O or S;</p>	<p><b>USES</b>          (I) are pesticides with insecticidal, nematocidal and acaricidal activities. Application rates are 0.01-50, pref. 0.1-10 lbs./acre.          In tests activity is shown variously against corn leaf aphid, boll weevil and southern corn rootworm at 1000 ppm, tobacco budworm at 6000 ppm and root knot nematode at 50 ppm.</p> <p><b>SPECIFICALLY CLAIMED</b>          10 cpds. (I), e.g. O,O-diethyl S-((4-oxo-4H-pyrido(1,2-a)pyrimidin-2-yl)methyl) phosphorodithioate (1a) and S-((6-chloro-5-oxo-5H-thiazolo (3,2-a)pyrimidin-7-yl)methyl) O,O-dimethyl phosphorodithioate.</p> <p><b>PREPARATION</b></p> <div style="text-align: center;">  </div> <p>EP--91700-A+</p>

Figure 1. Part of the Basic Abstract for a patent specification prepared by Derwent Publications Ltd. showing the generic structure description.

INPUT 91700 SD



R62 = ALKYL <1-3> / PHENYL ;

R1 = H ;

R2 = H / HALOGEN ;

R1+R62 = SD

- C = C - C = C -

OSB (CL/METHYL) /

[3/2] SD

- 1S - 2C = 3C -

[3/1] ;

R63 = O / S ;

R3 = ALKYL <1-3>;

R4 = ALKOXY <1-3> / THIO SB PROPYL ;

IF R62 = PHENYL THEN

THEN RESTRICT R3 = METHYL AND R4 = MeO.

Figure 2. GENSAI description of the generic structure shown in Figure 1, which illustrates a number of the features of GENSAI. All the variable groups have names of the form Rn, and where it has not been possible to use the original name, new names have been chosen to avoid confusion (63 is the maximum number of R groups permitted). Variables may be defined both separately and in combination, as with R1 and R62, and *position sets* (in square brackets) are used to indicate the orientation of the ethenediylthio group. IF and RESTRICT statements are used to impose special restrictions on the definitions of R62, R3, and R4.

storage medium but rather as the representation from which a range of degenerate representations can be produced on which faster, though less accurate, searches can be based. It is available for atom-by-atom search at the ultimate level of accuracy, should this be required.

The second and equally crucial requirement is the question of the searchable representations that can be derived from the

Table I. Structural Parameters for Generic Radical Groups

parameter	generic term	parameter	generic term
A	total atom count	RC	ring count
C	carbon count	RN	ring atom count
T	acyclic ternary	RS	ring substitution count
	branch count	RF	ring fusion count
Q	acrylic quaternary	RA	normalized ring count
	branch count	RZ	ring heteroatom count
E	alkene unsaturation count	Z	total heteroatom count
Y	alkyne unsaturation count		

Table II. Additional Structural Parameters Recommended for Generic Radical Groups

parameter	generic term	parameter	generic term
B	total acyclic branch count	ZO	total oxygen count
U	total unsaturation count	ZS	total sulfur count
		L	connectivity count
CZ	chain heteroatom count	V	valence(s) of connection bond(s)
		P	connection(s) via chain/ring atom(s)
ZN	total nitrogen count		

ECTR, so that a range of search algorithms operating at varying levels of explicitness, effectiveness, and speed can be applied. It is envisaged that the quality of the search mechanisms will improve over a period of some years from the initial feasibility demonstration, as experience is gained and as increasingly powerful and flexible search methods are developed. User aids at the search interface will also be required to assist in query development.

The searches to be supported include those in which queries comprise specific structures, for which inclusion as a member of a generic class should be the criterion for retrieval, generic structures, for which an overlap of one or more structures between the query and a file structure should be determining, and substructures, for which inclusion within a generic structure is the criterion. Fisanick has examined these requirements in some detail.<sup>56</sup>

These capabilities are not readily attained; the three query types pose difficulties that are in approximate order of their citation. The problems are of two kinds. Both file structures and queries of the second and third types may include generic radical terms in overlapping or nonoverlapping positions. Hence, the search routines must support both the generation

of characteristics that typify the structural features of generic radicals and their specific environments in a searchable representation and must also permit the recognition of correspondences between generic and specific radicals in query and file structure or vice versa.

In greater detail, the strategy adopted in the first respect has been to build on the earlier Sheffield work on the automatic generation (from specific connection tables) of highly discriminant atom- and bond-centered fragments,<sup>57</sup> work that has since been extended and implemented by the Swiss BASIC group<sup>58</sup> and adopted by Chemical Abstracts Service as the basis for the CAS Online screening search.<sup>59</sup> The strategy and the second element above, the recognition of correspondence between specific and generic radicals, have both called for the development of a novel chemical grammar<sup>50</sup> along lines earlier identified by Whitlock.<sup>60</sup>

The chemical grammar, TOPOGRAM, is applied in two senses. First, a rule-based production system, driven by the generic radical terms and their associated parameters, is used to generate the atom- and bond-centered screens that would characterize the generic radicals if all of the specific radicals of the class were to be generated. The screens so generated cover both the radicals and their specific or generic environments. Second, TOPOGRAM can be applied in a recognitive, rather than in a generative, sense. In this application, TOPOGRAM is used to test for membership of a substituent group in the query structure within the class of radicals described by a generic radical term in the file structure or, indeed, to test for common membership between two generic radical terms. This facility is most appropriately used in search algorithms subsequent to the screen search, in which correspondences between larger environments are sought.

TOPOGRAM has still to achieve the widest generality; one limitation, for instance, is that its screens characterize only ortho- and spiro-fused rings, and it must still be extended to characterize other ring-fusion modes, such as peri fusions and bridged systems. Again, its recognitive capabilities are still under development.

In the generic structure retrieval system currently under construction, the screens (presently a subset of those used in CAS Online) are generated in such a way that a distinction is made between those that occur wholly within the invariant part of a generic structure and those that occur partly or wholly in the variable parts, thus providing differentiation both in representation and in search.

To complement this fast serial search, other methods are under development, including atom-by-atom search routines into which the extended recognitive capabilities of TOPOGRAM will be included in time and a relaxation search method due to von Scholley,<sup>61</sup> which lies between the screening and atom-by-atom search in respect of its computational requirements. This technique has the attractive feature that it can be implemented as a parallel algorithm on suitable multiprocessor hardware (reflecting the growing interest in multiprocessors for both text<sup>62</sup> and chemical structure search<sup>63</sup>), although for the moment it is used only in a serial fashion.

While substantial problems of representing structural features of generic structures still await solution, the achievements of the Sheffield project are already substantial and hold much promise for the future for search systems for generic structures that parallel the capabilities of current specific structure retrieval systems.

#### CURRENT RELATED WORK

Recent reports from Japan detail further research on representing generic structures for search. Kudo and Chihara,<sup>64</sup> working with a small data base of structures from the Japanese Gazette List of Existing Chemical Substances, instance

searches on structures where nomenclatural terms lack locational specificity, as with "dichloroxylylene" and even alkane (C 10-29), while Nakayama and Fujiwara<sup>65</sup> instance the application of the block-cutpoint tree (BCT) to the factoring of generic expressions with logical conditions into simpler sets of expressions.

The principal services have announced their intention of implementing generic structure search systems; Chemical Abstracts Service announced a generic structure system for release in 1987, without giving detail.<sup>66</sup> For CAS, this would be a novel departure, in that their practice has thus far been to consider only the new chemistry in patents, including specific substances prepared and characterized. Again, Derwent Publications Ltd. recently announced an agreement to collaborate with the French Patent Office in the development of a generic structure search system by Télésystèmes.<sup>67</sup>

The coming years thus appear to promise as much activity in regard to generic structures as the 1970s and early 1980s saw in respect of specific substance search systems.

#### ACKNOWLEDGMENT

We gratefully acknowledge current and recent funds in support of the generic chemical structures work at Sheffield from the British Library Research and Development Department, Chemical Abstracts Service, Derwent Publications Ltd., and Internationale Dokumentationsgesellschaft für Chemie, m.b.H. We also thank J. P. Alexandrou, S. Ash, L. Carruthers, G. M. Downs, V. J. Gillet, R. M. Roscoe, and Dr. A. von Scholley, who have contributed to the research work, and the many others, both within and outside the University, with whom we have had helpful discussions.

#### REFERENCES AND NOTES

- (1) Perkin, William Henry Br. Patent 1984.
- (2) Singer, C.; Holmyard, E. J.; Hall, A. R.; Williams, T. I., Eds. "A History of Technology"; University Press: Oxford, England, 1958; pp 267-279.
- (3) Simmons, E. S. "Patent Family Databases". *Database* **1985**, *8*, 49-55.
- (4) Johns, T. M.; Adrus, W. G.; De Voe, S.; Myers, J.; Smith, R. G.; Uhler, O. "An Examination of the Leading Patents Equivalent Services". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 241-246.
- (5) "The Scientific and Technical Information contained in Patent Specifications"; Office for Scientific and Technical Information: London, 1973; OSTI Report 5177.
- (6) Yorke, B. A. "Pharmaceutical Patent Protection". *Med. Res. Rev.* **1984**, *4* (1), 25-46.
- (7) Jackson, S. E. "Experiences of a Patent Searcher". In "Computer Handling of Generic Chemical Structures"; Barnard, J. M., Ed.; Gower: Aldershot, 1984; pp 30-37.
- (8) Rosa, M. C. "Outline of Practice Relative to 'Markush' Claims". *J. Pat. Off. Soc.* **1952**, *34*, 324-325.
- (9) Valance, E. H. "Understanding the Markush Claim in Chemical Patents". *J. Chem. Doc.* **1961**, *1* (2), 87-92.
- (10) Bouman, H. "Too Prolific Markush". *J. Doc.* **1970**, *26* (2), 161-163.
- (11) Gordon, J. E.; Brockwell, J. C. "Chemical Inference. 1. Formalization of the Language of Organic Chemistry: Generic Structural Formulas". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 117-134.
- (12) Gordon, J. E. "Chemical Inference. 2. Formalization of the Language of Organic Chemistry: Generic Systematic Nomenclature". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 81-92.
- (13) Hyams, M. "Multilevel Retrieval of New Patent Compounds". In "Computer Handling of Generic Chemical Structures"; Barnard, J. M., Ed.; Gower: Aldershot, 1984; pp 202-217.
- (14) Huther, B. R. "Automating for United States Patent and Trademark Office: a Plan for the 1990's". *World Pat. Inf.* **1983**, *5* (1), 10-14.
- (15) Terapane, J. F.; Wolfe, L. A. "Chemical Database and Data Access Standards for the Patent Search File". In "Computer Handling of Generic Chemical Structures"; Barnard, J. M., Ed.; Gower: Aldershot, 1984; pp 179-201.
- (16) Smith, R. G.; Anderson, L. P.; Jackson, S. K. "Online Retrieval of Chemical Patent Information. An Overview and Comparison of Three Major Files". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 148-157.
- (17) Kaback, S. M. "Retrieving Patent Information Online". *Online (Weston, Conn.)* **1978**, *2*, 16-25.
- (18) Kaback, S. M. "Online Patent Searching: the Realities". *Online (Weston, Conn.)* **1983**, *7*, 22-31.
- (19) Kaback, S. M. "What's in a Patent? Information. But Can I Find It?" *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 159-163.
- (20) Suhr, C.; Harsdorf, E. von; 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"; Barnard, J. M., Ed.; Gower: Aldershot, 1984; pp 96-105.
- (21) Hyams, M. "Foreign Patents Documentation". *J. Chem. Doc.* **1966**, 6 (2), 101-123.
  - (22) Hyams, M. "Chemical Patents Information". *Chem. Br.* **1968**, 6, 416-420.
  - (23) Kaback, S. M. "A User's Experience with the Derwent Patent Files". *J. Chem. Inf. Comput. Sci.* **1977**, 17 (3), 143-148.
  - (24) Kaback, S. M. "Chemical Structure Searching in Derwent's World Patent Index". *J. Chem. Inf. Comput. Sci.* **1980**, 20 (1), 1-6.
  - (25) Kaback, S. M. "Derwent Search Aids". *Database* **1982**, 5 (3), 19-21.
  - (26) Norton, P. "Central Patents Index as a Source of Information for the Pharmaceutical Chemist". *Drug Inf. J.* **1982**, 208-215.
  - (27) Simmons, E. S. "Central Patents Index Chemical Code: A User's Viewpoint". *J. Chem. Inf. Comput. Sci.* **1984**, 24, 10-15.
  - (28) Fugmann, R. "The IDC System". In "Chemical Information Systems"; Ash, J. E.; Hyde, E., Eds.; Horwood: Chichester, 1975.
  - (29) Rössler, S.; Kolb, A. "The GREMAS System, an Integral Part of the IDC System for Chemical Documentation". *J. Chem. Doc.* **1970**, 10, 128-134.
  - (30) Kaback, S. M. "The API Chemical Indexing System". In "Computer Handling of Generic Chemical Structures"; Barnard, J. M., Ed.; Gower: Aldershot, 1984; pp 38-48.
  - (31) Kaback, S. M. "The IFI Plenum Chemical Indexing System". In "Computer Handling of Generic Chemical Structures"; Barnard, J. M., Ed.; Gower: Aldershot, 1984; pp 49-65.
  - (32) Balent, M. Z.; Emberger, J. M. "A Unique Chemical Fragmentation System for Indexing Patent Literature". *J. Chem. Inf. Comput. Sci.* **1975**, 15, 100-104.
  - (33) Ash, J.; Chubb, P.; Ward, S.; Welford, S.; Willett, P. "Communication, Storage and Retrieval of Chemical Information". Horwood: Chichester, 1985; pp 182-202.
  - (34) Silk, J. A. "Present and Future Prospects for Searching the Journal and Patent Literature". *J. Chem. Inf. Comput. Sci.* **1979**, 19, 195-198.
  - (35) Ray, L. C.; Kirsch, R. A. "Finding Chemical Records by Digital Computers". *Science (Washington, D.C.)* **1957**, 126, 814.
  - (36) Opler, A.; Baird, N. "Display of Structural Formulas as Digital Computer Output". *Am. Doc.* **1959**, 10, 59-63.
  - (37) Leibowitz, J.; Frome, J.; Andrews, D. D. "Variable Scope Search System: VS3". In "Proceedings of the International Conference on Scientific Information", Washington, DC, 1958; NAS/NRC: Washington, DC, 1959; Vol. II, pp 1117-1142.
  - (38) Frome, J.; O'Day, P. T. "A General Chemical Compound Code Sheet Format". *J. Chem. Doc.* **1964**, 4, 33-45.
  - (39) Koller, H. R.; Marden, E.; Pfeffer, H. "The HAYSTAC System, Past, Present and Future". In "Proceedings of the International Conference on Scientific Information", Washington DC, 1958; NAS/NRC: Washington, DC, 1959; Vol. II, pp 1143-1179.
  - (40) Hayward, H. W.; Tauber, S. J. "The HAYSTAC Experiment". In "Proceedings of the 5th Annual Meeting of the Committee for International Cooperation in Information Retrieval among Examining Patent Offices", London, 1965; Thompson: Washington, DC, 1966; pp 337-350.
  - (41) Tauber, S. J.; Bolotsky, G. R.; Chodos, E.; Fraction, G. F.; de Maine, P. A. D.; Friedman, H. J.; Kirby, C. L.; Marron, B. A.; Springer, G. K.; Walker, J. C. "Developing Computer Programs for Searching Specific and Generic Structures, Including the Formatting of Markush Structures". In "Progress in Techniques for Manipulating and Organising Chemical Information". National Bureau of Standards: Washington, DC, 1967; Report 9587, January 1967, pp 23-34.
  - (42) 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.; Feldmann, R.; Hyde, E., Eds.; Wiley: New York, 1974; pp 105-122.
  - (43) Meyer, E.; Schilling, P.; Sens, E. "Experiences with Input, Translation and Search in Files Containing Markush Formulae". In "Computer Handling of Generic Chemical Structures"; Barnard, J. M., Ed.; Gower: Aldershot, 1984; pp 83-95.
  - (44) Silk, J. "A Notation-Based Fragment Code for Chemical Patents". *J. Chem. Doc.* **1968**, 8, 161-165.
  - (45) Sneed, H. M. S.; Turnipseed, J. H.; Turpin, R. A. "A Line Formula Notation for Markush Structures". *J. Chem. Doc.* **1968**, 8, 173-178.
  - (46) Jackson, F. T. "Markush Structures". In "Proceedings of the CNA (UK) Seminar on Integrated Data Bases for Chemical Systems", University of Kent, Canterbury, April 1979; Chemical Notation Association (UK): London, 1981; pp 134-157.
  - (47) Krishnamurthy, E. V.; Lynch, M. F. "Analysis and Coding of Generic Chemical Formulae in Chemical Patents". *J. Inf. Sci.* **1981**, 3, 75-79.
  - (48) 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, 148-150.
  - (49) 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.
  - (50) 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.
  - (51) 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 (ECTR) for Generic Structures". *J. Chem. Inf. Comput. Sci.* **1982**, 22, 160-164.
  - (52) 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.
  - (53) 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.
  - (54) Welford, S. M.; Lynch, M. F.; Barnard, J. M. "Towards Simplified Access to Chemical Structure Information in the Patent Literature". *J. Inf. Sci.* **1983**, 6, 3-10.
  - (55) Welford, S. M.; Ash, S.; Barnard, J. M.; Carruthers, L.; Lynch, M. F.; Scholley, A. von. "The Sheffield University Generic Chemical Structures Research Project". In "Computer Handling of Generic Chemical Structures"; Barnard, J. M., Ed.; Gower: Aldershot, 1984; pp 130-158.
  - (56) Fisanick, W. "Requirements for a System for Storage and Search of Markush Structures". In "Computer Handling of Generic Chemical Structures"; Barnard, J. M., Ed.; Gower: Aldershot, 1984.
  - (57) Adamson, G. W.; Cowell, J.; Lynch, M. F.; McLure, A. H. W.; Town, W. G. "Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files". *J. Chem. Doc.* **1973**, 13, 153-157.
  - (58) Graf, W.; Kaendl, H.; Kniess, H. K.; Schmidt, B.; Warszawski, R. "Substructure Retrieval by Means of the BASIC Fragment Search Dictionary Based on the Chemical Abstracts Service Registry III System". *J. Chem. Inf. Comput. Sci.* **1979**, 19, 51-55.
  - (59) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. "The CAS Online Search System. I. General System Design and Selection, Generation and Use of Search Screens". *J. Chem. Inf. Comput. Sci.* **1983**, 23, 93-102.
  - (60) Whitlock, H. W. "An Organic Chemist's View of Formal Languages". *ACS Symp. Ser.* **1977**, No. 61, 60-80.
  - (61) Scholley, A. von. "A Relaxation Algorithm for Generic Chemical Structure Screening". *J. Chem. Inf. Comput. Sci.* **1984**, 24, 235-241.
  - (62) Pogue, C.; Willett, P. "An Evaluation of Document Retrieval from Serial Files Using the ICL Distributed Array Processor". *Online Rev.* **1984**, 8, 569-584.
  - (63) Wipke, W. T.; Rogers, D. "Rapid Subgraph Search Using Parallelism". *J. Chem. Inf. Comput. Sci.* **1984**, 24, 255-262.
  - (64) Kudo, Y.; Chihara, H. "Chemical Substance Retrieval System for Searching Generic Representations. 1. A Prototype System for the Gazetted List of Existing Chemical Substances of Japan". *J. Chem. Inf. Comput. Sci.* **1983**, 23, 109-117.
  - (65) 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.
  - (66) *CAS Rep.* **1984**, 17, 8-9.
  - (67) Emard, J.-P. "Derwent, INPI and Télé systèmes Sign Agreement". *Online (Weston, Conn.)* **1985**, 9 (3), 14.