of this paper. I am also grateful to Drs. R. Fugman, A. L. Goodson, K. L. Loening, M. F. Lynch, and William Wiswesser for constructive remarks and criticism through private discussions and correspondence.

REFERENCES AND NOTES

(1) Fletcher, J. H.; Dermer, O. C.; Fox, R. B. "Nomenclature of Organic Compounds"; American Chemical Society: Washington, DC, 1974.
(2) Goodson, A. L. "Graph-Based Chemical Nomenclature. 1. Historical Background and Discussion. 2. Incorporation of Graph-Theoretical Principles into Taylor's Nomenclature Proposal". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 167-176.
(3) Smith, E. G. "The Wiswesser Line Formula Chemical Notation"; McGraw Hill: New York, 1968.
(4) Krishnamurthy, E. V.; Sankar, P. V.; Krishnan, S. "ALWIN-Algorithmic Wiswesser Notation System for Organic Compounds". *J. Chem. Doc.* **1974**, *14*, 130-141.
(5) Krishnan, S.; Krishnamurthy, E. V. "Compact Grammar for ALWIN Using Morgan name". *Inf. Process. Manage.* **1976**, *12*, 19-34.
(6) Backus, J. W. "The Syntax and Semantics of the Proposed International Algebraic language of the Zurich ACM-GAMM Conference". Information Processing, Proceedings of ICIP Paris, UNESCO, Paris, pp 125-132.
(7) Aho, A. V.; Ullman, J. D. "Principles of Compiler Design"; Addison-Wesley: Reading, MA, 1977.

Computer Storage and Retrieval of Generic Structures in Chemical Patents. 4. An Extended Connection Table Representation for Generic Structures

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

Department of Information Studies, University of Sheffield, Sheffield S10 2TN, United Kingdom

Received April 5, 1982

A data structure for the unambiguous representation of generic structures at the machine level is described. It is designed for automatic generation from structures encoded in the formal language GENSAL and is based on connection tables. Its relationship with other forms of representation is discussed.

INTRODUCTION

Paper 2 in this series[1] described a formal language, GENSAL, which has been designed for the encoding of generic, or Markush, structures in a form which is intelligible to a chemist or patent agent and also amenable to automatic analysis by computer. It was suggested that GENSAL is analogous to a high-level programming language and that the program which analyzes it can be thought of as equivalent to a compiler. To extend this analogy further one can compare the internal representation generated by the GENSAL interpreter program with the object code produced by a programming language compiler, and though unlike the object code for a programming language, the internal representation is a machine-level data structure rather than a set of machine-level instructions.

In a generic structure information system, this internal representation can be used to generate fragments for use in searching or used directly for atom-by-atom tracing in the final stage of a search. In order to enable it to perform these functions satisfactorily, and yet remain in a form which can easily be generated from GENSAL input, we have incorporated a number of features into its design, and these will be described in this paper. More detailed considerations of its use in fragment generation and searching will be the subject of future papers in this series.

REQUIREMENTS FOR THE REPRESENTATION

The first paper in this series[2] discussed the need for a full and unambiguous description of the generic structure, from which fragment screen descriptors of various types could be generated algorithmically, and the reasons for the selection of connection tables as the appropriate basis for this representation.

The purpose of the representation described here is not to store explicitly all the possible specific structures covered by a given generic structure but rather to contain sufficient information for exhaustive generation of all the specific structures to be possible, even though in most cases such an operation would be pointless, as well as computationally unfeasible where the number of specific structures covered in large, or even infinite.

Since the representation is to be built up from a generic structure input to the computer in GENSAL, the conversion problems will be greatly simplified if certain features of the representation mirror features of GENSAL. In particular, as the syntax for the definition of *substituents* in GENSAL[3] is essentially recursive, the structure of the internal representation should be recursive also.

GENSAL views a generic structure as consisting of a (possibly vestigial) constant part, to which are attached variable parts which can vary in their chemical nature, position of attachment, and multiplicity of occurrence and which may themselves be further substituted by other constant and variable parts down to any level. At each level, certain of the values for the variable parts may be alternative or additional to each other in complex nested Boolean relationships. This suggests two principal components for the internal representation, one containing information about the chemical nature of the constant and variable parts and the other containing information about the way in which they are connected together in terms of positions and frequencies of occurrence and the Boolean relationships between them. The successive levels of further substitution imply a hierarchical relationship between the different parts of the structure, though the exact nature of the hierarchy depends on the way in which the GENSAL description of the structure was constructed, which is to a certain extent arbitrary. It is also possible for the hierarchy to "loop back" to a higher level, in which case a recursive definition of a substituent (i.e., one in which the substituent is defined in terms of itself) will appear in the GENSAL. In this case there is no lowest level of substitution, and the structure in question is a polymer.

Together the two components of the internal representation can be considered as forming a topological graph, the chemical nature of the various parts of the generic structure being

© 1982 American Chemical Society

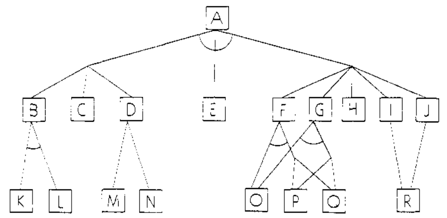
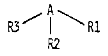


Figure 1. Diagrammatic representation of the basic structure of the ECTR showing the Child Gates. Each box represents a partial structure, and the lines represent Child Gates. Each hierarchical level of substitution is shown as a separate row of partial structures. Lines meeting at a point connect together partial structures which are alternative to each other (OR relationship), and lines meeting at a point that are linked together by an arc connect partial structures which are additional to each other (AND relationship). The GENSAL statements corresponding to this ECTR are shown in Figure 2.

```
INPUT  12345

SD
```



```
R1 =  B  SB  K  ANDBY  L    /   C   /  D  SB  ( M / N ) ;

R2 =  E ;

R3 =  ( F / G )  SB  O  ANDBY  ( P / Q )   /
                 H  /  ( I / J )  SB  R  .
```

Figure 2. GENSAL statements corresponding to the ECTR illustrated in Figure 1.

represented in the nodes of the graph and the information about their connections and relationships in its edges. Since information on the chemical nature of each part is predominantly based on conventional connection tables, the whole is a sort of super connection table, or connection table of connection tables, and is referred to as an Extended Connection Table Representation (ECTR). Within the ECTR each node is called a partial structure (PS) and each edge a gate. The gates are divided into "Child Gates" and "Parent Gates", according to which direction in the hierarchy they point: the graph is thus a *directed* one. The overall layout of the ECTR for a generic structure, showing the PSs and the Child Gates, is shown in Figure 1 (see Figure 2 for the GENSAL statements).

The entire ECTR is held in the main computer memory during its generation because, as further parts of the structure are defined during the course of the GENSAL sentence, it is frequently necessary to refer back to previously defined parts. Similarly, as fragments are generated or an atom-by-atom search performed, it is necessary to trace from one PS to another. The ECTR has been implemented by using data types of the programming language Pascal, in which the interpreter program is written.

PARTIAL STRUCTURE RECORD

From the syntax for a GENSAL *substituent value* it is possible to recognize four different types of partial structure that may be found in generic structures, and each of these requires a different representation in the ECTR.

"Specific" Partial Structures. These correspond to a single fully defined structural entity and are the only type of PS that may be represented by a connection table. They appear in GENSAL substituent values as structure diagrams or as specific nomenclatural terms such as "phenyl" or "nitro" (which the GENSAL interpreter program translates into connection tables via a dictionary of standard nomenclatural terms).

"Generic" Partial Structures. These appear in GENSAL substituent values as homologous series terms such as "alkyl" with associated parameter lists. They are shown in PS records as expanded parameter lists, including those parameters implied by the homologous series term itself as well as those given explicitly (for example, the term "alkenyl" implies at least one double bond which could be indicated by the parameter E⟨1–⟩). This type of PS is handled for fragment generation and searching by using chemical grammars.[4]

Table I. Partial Structure Record

Child Gate			
Parent Gate			
Specific	Generic	Unknown	Other
Connection Table	Parameter List	-	Character String

Table II. Connection Table Row

Atom Type	Substituent Name
	Substituent Values
Charge	
Number of Hydrogens	
Six Congeners	

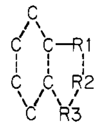
"Unknown" Partial Structures. These appear in GENSAL substituent values as a question mark and are most usually found in expressions such as "optionally substituted phenyl", which is shown in GENSAL by "phenyl osb ?". Clearly, no further information can be stored about their chemical nature, and search algorithms should allow them to be matched against any structural entity.

"Other" Partial Structures. These cannot be directly associated with any particular structural characteristics and include expressions such as "electron-withdrawing group" or "easily hydrolyzed group". They are shown in PS records as a character string, taken from the "other term" in the GENSAL substituent value, and could be used for some sort of text-based searching. Table I summarizes the information given in a partial structure.

CONNECTION TABLE FORMAT

The connection table used to represent specific PSs is a simple redundant one, each row representing one node which may be either an atom (in which case the atom type is recorded as a two-letter symbol) or a GENSAL substituent (in which case its name—the "R1", "R2", etc. of GENSAL—is recorded along with the values it can take in the same format as a Child Gate). The record structure for a connection table row is shown in Table II.

Normally, substituents attached to a specific PS are not explicitly included in the connection table as information about the atoms to which they are connected is stored in the Child Gates. It is only when there is a chain (cyclic or acyclic) of such substituents connected together, as shown below, that it is necessary in order to indicate the order in which they are connected to each other.



The number of attached hydrogen atoms is recorded for each row in order to permit the determination of the positions available for substitution in each PS. For commonly occurring atoms the interpreter program is able to calculate the number of hydrogens from the possible valencies.

Up to six congeners are possible for each row, this being a restriction imposed by the Feldmann structure diagram graphics system used,[5] and the record structure for each is shown in Table III. Other graphics systems might relax the

Table III. Congener Record

None	Fraternal	Filial	Parental
-	Row number of connected atom or "NOTFIXED" for variable-position connection	Substituent	-
Bond Order			

Table IV. Homologous Series Term Parameters

Carbon Count
Ternary Branch Points
Quaternary Branch Points
Double Bonds
Triple Bonds
Number of Rings present
Number of Atoms in Rings
Number of Substitutions on Rings
Number of Aromatic Rings
Number of Heteroatoms Present

Table V. Item in Combination Bar of Child Gate

Positions in Parent PS	
Frequency of Occurrence	
Bottombar	Not Bottombar
Positions in Child PS	Pointer to Alternative Bar (next layer)
Bond Order	
Pointer to Child PS	
Pointer to next item in Combination Bar list	

Table VI. Item in Alternative Bar of Child Gate

Pointer to Combination Bar (next layer)
Pointer to next item in Alternative Bar list

limitation. For each are recorded a bond order (again, these are taken from the Feldmann system) and information about the nature of the connected node. "Fraternal" connections are those to other rows in the same PS: the relevant row number is recorded. "Filial" connections are those to other PSs "lower down" in the ECTR, and a substituent name is recorded, though this is only used for certain housekeeping operations during the setting up of the ECTR; details of the connection are given in the Child Gate. "Parental" connections are those to other PSs "higher up" in the ECTR, and details are given in the Parent Gate.

An arbitrary limit of 32 rows is set for each connection table and is thus the maximum number of nonhydrogen atoms permitted in a structure diagram in GENSAL. However, because the splitting of a generic structure into separate PSs is to a certain extent arbitrary, a large structure diagram can always be divided into two or more smaller ones.

PARAMETER LIST FORMAT

The third paper in this series[4] described the manner in which the *parameters* applied to a homologous series term in a GENSAL sentence could be used to apply constraints to the chemical grammars used for generation and/or recognition of the members of the homologous series. A set of standard parameter identifiers for constraints on such features as atom count, branch points, and unsaturations was given, and the second paper of the series[1] described the possibility of using nonstandard parameters to indicate interruptions in a chain or substitutions on it.

Extension of the chemical grammars to include certain types of cyclic systems has led to the definition of further standard parameters, and the set currently in use is shown in Table IV. These extensions and algorithms for the automatic generation of fragment descriptors from the constraints on a particular homologous series term, without the need for exhaustive generation of all the possible structures covered, will be described in a future publication.

The full set of parameters with their values is sufficient, when used to constrain the chemical grammars, to define completely all the possible structures covered. Consequently, the PS record for the generic type of PS can consist simply of a list of parameter values (as integer ranges) for all the standard parameters. The nonstandard parameters are treated as substitutions on and within the generic PS, and information about them is given in Child Gates, as described below. However, when fragments are generated or paths traced within the ECTR, the information about "Children" of generic PSs will be used to apply constraints to the chemical grammars.

CHILD GATE FORMAT

Child Gates indicate the connections from one PS (called the Parent PS) to those lower down in the hierarchy to which it is connected. There may be connections to several Child PSs, which can be additional or alternative to each other. Each Child Gate therefore describes a "one-to-many" relationship, though over the ECTR as a whole the Child Gates between successive levels of the hierarchy describe a "many-to-many" relationship, as can be seen from Figure 1.

Child Gates are arranged in a series of alternating "Bars" of two mutually recursive types: the first, called a "Combination Bar", gives information about Child PSs which are additional to each other (i.e., which all occur combined together on the Parent PS), and the other, called an "Alternative Bar", gives information about Child PSs which are alternative to each other (i.e., only one of which occurs on the Parent PS). These two types of Bar appear in alternating layers to any necessary depth, and the number of layers will depend on the amount of bracketing (indicating recursive *substituent definitions*) in GENSAL. Both types of bar are constructed as linked lists of items, the items in a Combination Bar being those which are additional to each other and the items in an Alternative Bar being those which are alternative to each other.

Combination Bars. The record structure of an item in a Combination Bar is shown in Table V. If alternatives are possible for an item in a Combination Bar, a pointer is given to an Alternative Bar in the next layer of the gate. If no such alternatives are possible, a pointer is given to the appropriate Child PS record, along with information about the positions in the Child PS at which the attachment may be made and the order of the connecting bond, and the Bar is said to be a "Bottombar". Both Bottombar and non-Bottombar Combination Bar items contain information about the positions in the Parent PS at which the Child PS may be attached and the number of times that it can occur in these positions. The positions are taken from an explicit GENSAL position set (if present) or calculated from those positions available for substitution. The frequency may be given in a GENSAL *selector* or in the definition of a *multiplier*, or, if the Child has been specified in a parameter list for a homologous series term from the values given for that parameter, it may be calculated from the number of positions available. In the case of non-Bottombar items this position and frequency information applies to all the alternatives specified in the Alternative Bar pointed to.

Alternative Bars. These have a much simpler structure than Combination Bars, and the record structure for an Alternative Bar item is shown in Table VI. All the information about each alternative in the list is given in the Combination Bar pointed to.

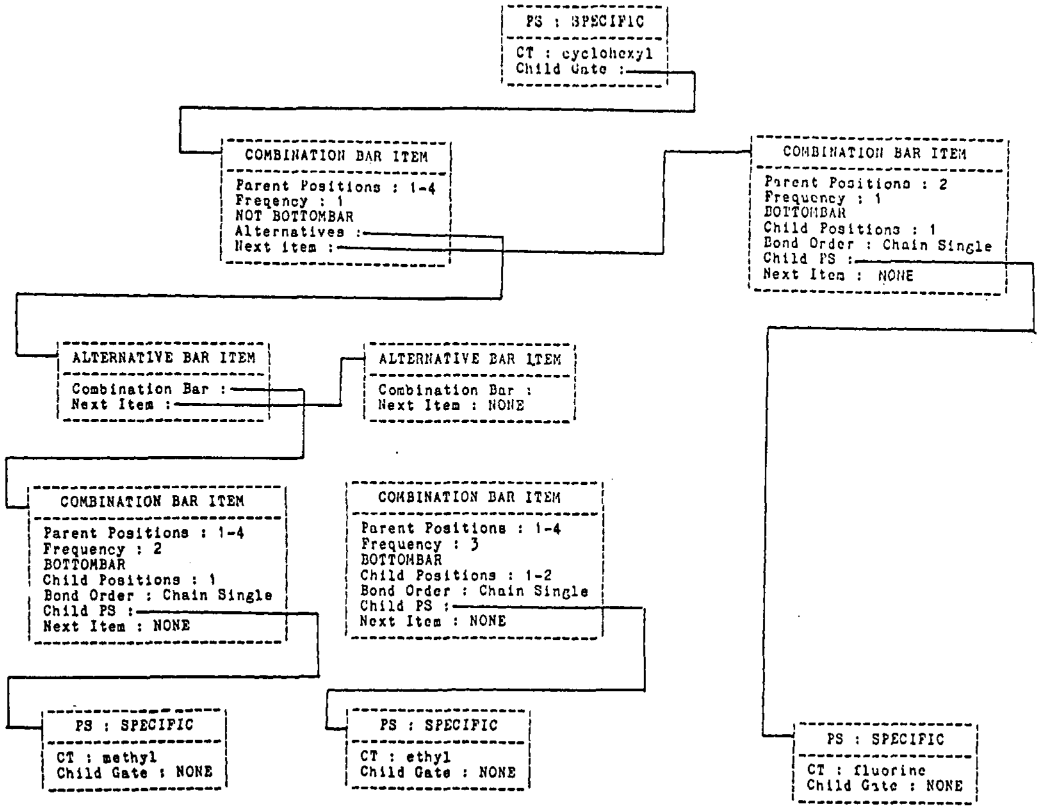


Figure 3. Diagrammatic representation of the structure of the Child Gate corresponding to the GENSAL expression cyclohexyl SB [1–4] (⟨2⟩ methyl / ⟨3⟩ ethyl) ANDBY [2] F, which means that cyclohexyl is substituted in positions 1, 2, 3, and/or 4 by either two methyl groups or three ethyl groups and in addition to these by one fluorine in position 2.

Table VII. Item in Parent Gate

Positions in Child PS
Positions in Parent PS
Bond Order
Pointer to next item in Parent Gate

The "top" Bar of a Child Gate, which is the point at which it is accessed from the Parent PS, is always a Combination Bar, and the Child Gate field of a PS record (Table I) is therefore a pointer to a Combination Bar item. Figure 3 illustrates the internal structure of a single Child Gate for a moderately complicated GENSAL expression.

PARENT GATE FORMAT

The structure of Parent Gates is very much simpler than that of Child Gates, as none of the information on the Boolean relationships between the various Child PSs is stored in them. In fact, all the information contained in a Parent Gate is also contained in the corresponding Child Gates, and the purpose of Parent Gates is simply to allow path tracing within the ECTR to take place from a Child PS to a Parent PS as well as in the other direction; the redundancy of the information in the Parent Gates is compensated for by the substantial enhancements in path-tracing ability.

Like the two types of Bar in Child Gates, Parent Gates are implemented as a linked list of items, each item referring to a different possible Parent PS for the Child in question. The record structure is illustrated in Table VII. For each possible Parent PS, the possible positions for connection in both the Child and the Parent are given, along with a pointer to the Parent PS and the order of the connecting bond.

The Parent Gate field of a PS record gives a pointer to the first item in a linked list of Parent Gate items. Figure 4 illustrates the overall structure of the Parent Gates for the same generic structure as was used to illustrate Child Gates in Figure 1.

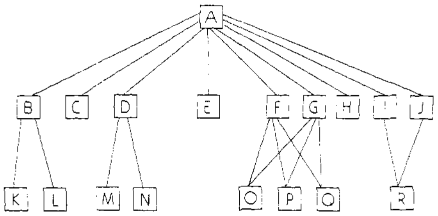


Figure 4. Diagrammatic representation of the ECTR for a generic structure showing the Parent Gates. The generic structure is the same as those shown in Figures 1 and 2.

REPRESENTATION OF CONDITIONS AND RESTRICTIONS

The ECTR described above makes no provision for incorporating the information given in GENSAL "IF" and "RESTRICT" statements nor for distinguishing between the five different assignment operators that can be used to indicate independent or nonindependent values for substituents or multipliers.

These features of GENSAL, which mirror many of the expressions found in chemical patent specifications, are used to limit the variety of possible specific compounds covered by a generic structure by restricting the co-occurrence of particular alternatives in substituent definitions etc. The present form of the ECTR may thus describe a greater variety of specific compounds than is actually warranted, and the limitations imposed by "IF" and "RESTRICT" statements could be implemented by indicating which of the possibilities in the ECTR should not co-occur. This might be achieved by applying some sort of selective "lock" to the gates, though the way in which this might be represented in the computer has yet to be determined.

COMPARISON OF ECTR WITH OTHER REPRESENTATIONS

Silk[6] has drawn attention to the similarity between a Markush structure and a nested Boolean expression and suggested that the Boolean relationships could be incorporated into a notation-based representation for generic structures. The ECTR also exploits this similarity with the successive layers of Bars in Child Gates representing the nested Boolean relationships, though the PSs are represented by connection tables rather than by notation strings.

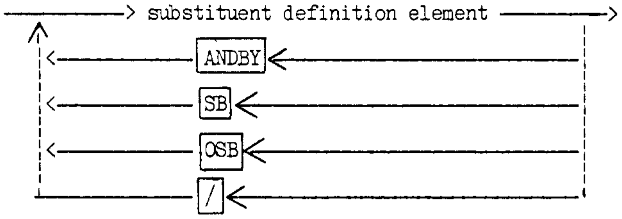
An approach much closer to that described here has been proposed by Fugmann et al.,[7] who suggested that the topological graphs used to represent concept relationships in the TOSAR system of the IDC might also be used to indicate the Boolean relationships present in a Markush formula, though warning that the amount of space required to store the graph could cause difficulty and that the machine time needed for searching would increase dramatically if many cycles were present.

The Chemical Abstracts Registry III System[8] employs a mechanism for compiling several partial connection tables to describe a larger structure, though in this instance it is used as a space-saving measure in the storage of specific structures having ring systems in common and not as a means of describing generic structures.

It has not so far been possible to evaluate the space requirements for the ECTR or the machine time needed for fragment generation or atom-by-atom searching. However, it is intended that the ECTR should only be retained for the purpose of fragment generation and not be stored permanently, and this is likely to ease the former problem at least. No arbitrary limits are applied to the size of ECTR that may be generated, and if it is required for atom-by-atom searching, it can always be regenerated from the stored GENSAL statements at the time that it is needed. It is recognized that atom-by-atom tracing in the ECTR will be very expensive in computer time (much more so than for such searching in specific structures), but it is hoped that it is only in very rare cases that it would be necessary. The extent to which it will be needed will be determined by the effectiveness of the search algorithms devised for use with fragment descriptors linked by their logical relationships, as outlined in the first paper of this series.[2]

APPENDIX I. MODIFICATIONS TO GENSAL SYNTAX

Substituent Definitions. Some minor weaknesses in the syntax of *substituent definitions* have been removed, though this does not significantly affect GENSAL as written. It has been found possible to do without the alternative substitution operator ("ORBY"), "/" being used instead, and a rigid operator precedence has been established, whereby "ANDBY" is evaluated first, followed by "SB" and "OSB" (which rank equally) and finally by "/", which is evaluated last. Expressions within brackets are evaluated before "ANDBY". The syntax diagram for substituent definition thus becomes:



Integer Ranges. The syntax of integer ranges has been simplified by the removal of the requirement for angle brackets around the range, which avoids the rather clumsy construction of *position sets* with both square and angle brackets, as in [⟨2–4⟩]. The angle brackets have, however, been retained for *selectors*.

ACKNOWLEDGMENT

Thanks are due to the staff of Sheffield University Computing Services, to Dr. P. Willett for helpful discussions, and to the British Library Research and Development Department for financial support.

REFERENCES AND NOTES

(1) 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.

(2) 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.

(3) Certain minor modifications to the syntax of GENSAL have been made since the paper describing it was published. These are described in Appendix I.

(4) 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.

(5) Feldmann, R. J.; Milne, G. W. A.; Heller, S. R.; Fein, A.; Miller, J. A.; Koch, B. "An Interactive Substructure Search System". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 157–163.

(6) Silk, J. A. "Present and Future Prospects for Structural Searching of the Journal and Patent Literature". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 195–198.

(7) Fugmann, R.; Nickelsen, H.; Nickelsen, I.; Winter, J. H. "Representation of Concept Relations Using the TOSAR System of the IDC. Treatise III on Information Retrieval Theory". *J. Am. Soc. Inf. Sci.* **1974**, *25* (5), 287–207.

(8) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Registry System. 1. General Design". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111–121.