dividual, since the proportion of reading done on personal time vs. company time is not known.)

time value, 2 h/month × \$30–50/h/professional =
\$60–100 saving/prof/month = \$720–1200/prof/year

Even at this conservative value level, the ratio of benefits to costs for the *Updates* is well over a factor of 10.

EXPERIMENT SUMMARY

Updates for Exxon have effectively filled a recognized gap in access to and awareness of current developments in technology pertinent to the needs of Exxon Chemical. The contents of the *Updates*, as currently prepared, are relevant to the needs of over 90% of the recipients and substantial numbers of original articles are requested on the basis of *Updates* abstracts.

An important finding is that in the absence of a targeted subject-oriented bulletin, such as *Updates*, a large number (over 30%) of users would attempt no current-awareness program at all, while others would request more literature services or attempt, at a cost in time, to find nearby library resources to fill their needs.

This experimental program has provided CAS with the opportunity to demonstrate its ability to produce timely and relevant information in a package designed for a geographically dispersed group of Exxon scientists and engineers with a common subject interest. The result for the Exxon staff has been access to information support where there was little on site. Savings to Exxon in time and other benefits are estimated to be substantial and sometimes directly measurable.

BEYOND PRESENT UPDATES

Variations to the service described in this paper are possible, and these could enable information base producers to provide the end users of information with potential results not unlike this experiment with Exxon. Some possibilities are: business information to support scientific investigations; services based on information needs stimulated by the regulatory actions of government agencies to satisfy the responsibilities of patent, engineering, health, personnel, and legal departments; services which are responsive to the interests of a targeted audience and combine information from a variety of bases.

The advent of powerful information-based production systems affords new opportunities for the technological community to request information such as found in individualized information services and for producers to satisfy these requests. Indeed, studies are currently underway with Exxon to develop the concepts utilized in the *Updates* into other fields of technology and business interests.

REFERENCES AND NOTES

(1) Blake, J. E.; Mathias, V. J.; Patton, J. "CA SELECTS–A Specialized Current Awareness Service", *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 187–190.

(2) Nie, Norman H., et al. "Statistical Package for the Social Sciences", 2nd ed. McGraw-Hill: New York; 1975.

(3) Weil, B. H. "Some Reader Reactions to Abstract-Bulletin Style", *J. Chem. Doc.* **1961**, *1*, 52–58.

Unique, Unambiguous Representation of Chemical Structures by Computerization of a Simple Notation

RÜDIGER WALENTOWSKI

Beilstein-Institut für Literatur der Organischen Chemie, Frankfurt am Main, Federal Republic of Germany

Received January 15, 1980

SNN (*S*tructure-*N*omenclature *N*otation) is based on the features of chemical structure and uses rules derived from conventional chemical nomenclature. The molecule is split into fragments by structure-determining vertices. These fragments are coded (C number and heteroatom symbols) and linked together by special signs. The computer assigns to each compound the BUS (*B*eilstein *U*nique *S*equence) number, according to which the compound can be assigned its proper place within the Beilstein System. Another canonical representation derived from SNN is the Fragment Connection Table (FCT), which requires significantly less machine storage than conventional atom connection tables.

INTRODUCTION

The best and least ambiguous form of representation of a chemical compound is its structural formula. Furthermore, a linear representation is desirable, briefly characterizing the molecule and making it easy to locate in registers. Nomenclature has up to now fulfilled this task, albeit in an ever-worsening manner, as common names are frequently not unique and sometimes ambiguous, and systematic names are often difficult to handle because of their length, and complicated rules are called for.

Several linear notations manage a fairly brief form of representation, but the rules are so complicated that general use is out of the question. The aim of this paper is to describe a method which combines well-known elements of nomenclature and well-known elements of graphic representation to a uniform linear notation. This notation must be readily understandable, easy to handle, and suitable for computerization.

The method is founded upon the pioneering work of F. K. Beilstein, universally recognized to be the first to organize organic chemical compounds into a reference book in a systematic manner.

Some Historical Background of the Beilstein System. In 1881 Friedrich Konrad Beilstein published the first edition of the handbook that is named after him. After the third edition the number of known chemical compounds rose to over 100 000, and it became clear that a special system needed to be developed in order to catalog logically the multitude of compounds. The Beilstein System in use to the present day was developed and tested between 1907 and 1912 by B. Prager, P. Jacobson, P. Schmidt, and D. Stern. Great care was taken that each compound would be unfailingly classified in its proper place.

0095-2338/80/1620-0181\$01.00/0

Strict handling of this principle caused the separation of chemically related compounds in some cases; e.g., succinic anhydride will not be found under succinic acid (category Acyclic compounds) but in the category Heterocyclic compounds. Experience has shown that even compounds with structures unknown when the system was first developed can be easily classified. The system is in every sense open to expansion and unlimited.

Fundamentals of the Beilstein System. The basis of the classification of an organic compound is its structure. The main characteristic is the arrangement of the carbon atoms. This arrangement can be acyclic (open carbon chains), isocyclic (rings with members consisting entirely of carbon atoms), or heterocyclic (including a ring with members consisting partly of carbon atoms and partly of other atoms). This then corresponds to the three main categories of the Beilstein System.

The replacement of a hydrogen atom in the basic skeleton by a univalent hydrogen-containing group (e.g., OH, NH_2) results in a compound with a defined chemical functional group. Basic skeletons with or without functional groups will be referred to as stem compounds. When two hydrogen atoms are substituted on a C atom, these are treated as an oxo function due to a formal hydrolysis. Analogously we obtain the acid function for substitution of three hydrogen atoms. Functional groups are listed in Table IV.

Functional derivatives of the stem compound are obtained when the hydrogen of a functional group is replaced by organic substituents. These derivatives are ordered in that sequence which arises when considering the substituent on its own (i.e., "split off" from the stem compound) and assigned its own system number (see below). The sequence then is determined first by the system number of the stem compound, and second by that of the substituent.

In the case of there being more than one cleavage point, that basic skeleton shall be the stem compound that is classified at the systematically latest point (Principle of Latest Possible Systematic Entry). This splitting at carbon atom/heteroatom bonds (not, however, within heterocyclic rings) constitutes the basic principle of the Beilstein System.

Further classification of stem compounds obtained in this way is based on structural characteristics as outlined below. System numbers were created to facilitate practical use. However, the Beilstein System number bears no direct correlation to the chemical structure, and the rules for a purposeful tracing of a compound are rather extensive. To date the system has been unable to find a wider application.

MODERN METHODS FOR REPRESENTING CHEMICAL STRUCTURES

Past decades saw an intensive search for new ways to represent the rapidly growing number of chemical compounds in such a way as to facilitate mechanical manipulation. Three basic possibilities became apparent: fragment codes, topological methods, and linear notations.

Fragment Codes. Structural fragments are prespecified and assigned code symbols. The GREMAS code[1] and the Ring Code[2,3] are fragment codes in use. Compounds with defined substructures can be obtained simply and with relatively little calculation time. However, compounds might also be obtained during the search which were not sought for (false drops). Furthermore, only a limited number of structural characteristics can be defined, and the individual compound is not described unequivocally.

Topological Methods. All atoms of a compound and their bonds to each other are represented in tabular form (connection tables). Mastering the problem of computer input (see below) has brought topological methods into wide use, e.g., with *Chemical Abstracts*[4] and the IDC (Internationale Dokumentationsgesellschaft für Chemie).[5] The somewhat voluminous topological tables were reduced to a less redundant form for the purpose of computer storage. These have to be reconverted, however, to a redundant form for the purposes of searching. To avoid having to go through every topological storage set, preselections are carried out. It is clear that the search involves a fair amount of expense.

Linear Notations. Atoms or groups of atoms and their structural environment are described by code symbols in strictly defined sequence. The best known notations are Dyson (IUPAC notation)[6] and Wiswesser (WLN).[7] A number of other codes and notations have also been suggested (e.g., ref 8–11). The principal advantage of a linear notation is a compact and unambiguous representation. However, substructures can often be found only via the detour of topological representation. In the ICI Crossbow system,[12] for instance, the WLN is converted to a connection table for this purpose.

A number of rules are necessary in order to reach a unique representation, and these frequently require expansion in the case of new, previously unknown, or unconsidered structures (see supplementary notes to WLN; cf. ref 13). As a linear notation can be somewhat impenetrable for the chemist, a second code is often employed showing more clearly the chemical relationships involved. The Index Chemicus Registry System, for instance, has additionally introduced the Ring Code to compensate for the drawbacks of WLN.[14]

Representation of Chemical Structure in the Beilstein System. The Beilstein System cannot be classified within the abovenamed methods of structure representation. The system number is first and foremost a classification according to ring type, function type, degree of unsaturation, number of C atoms, etc. It is perhaps less widely appreciated that this classification progresses until a compound is assigned a unique status, tracing back to further structural characteristics such as number of rings, size of rings, condensation points, branching, etc. Finally their arrangement and connection in the molecule is considered.

When carrying out a manual search, classes of compounds can be found as well as a definite compound by consistent application of the Beilstein rules. The unique location of a compound is laid down by the Beilstein volume number and the page number (possibly also the entry number). The system is superior to the linear notations used hitherto in that it reveals chemical relationships more clearly. Formal splitting takes place at the heteroatom positions (where the molecule is most reactive), so that derivatives remain mostly related. The aim of this paper is to show a way to utilize these advantages with the aid of a code. Knowledge of the Beilstein System is not required for the setting up of the code. The code should be suitable for mechanical processing.

METHODS OF COMPUTER INPUT OF CHEMICAL STRUCTURES

There are two basic methods:

(a) Graphical Input. The structural formula is drawn on a grid and the coordinates as well as types of atoms and bonds are entered into the computer; similarly, the structural formula is drawn on a graphic display via a special keyboard, and this information is topologically coded by the computer.

(b) Input of a Code. The structural formula is translated into a code and this is input to the computer.

The rate-determining factor in the operation is the correct representation of the structural formula in the first instance, coding in the second.

The first method, graphical input, calls for more apparatus. Several paths were explored.

In 1962, Meyer constructed a machine for reading formulas, whereby punch cards could be perforated at certain points with the aid of photocells.[15]

Ohnacker and Kalbfleisch[16] reported a method of drawing the structure formula on a grid and putting into a puncher the coordinates, types of atoms, and bonds. They were able to process 40 formulas per hour.

Corey and Wipke[17] and Koniver, Wiswesser, and Usdin[18] constructed an apparatus that uses a slate pencil to draw the formula onto a prepared tablet (rand tablet) which then appears immediately on a projection screen and is entered into the computer.

Neubert[19] developed a process where the structure formula is "drawn" on the screen with the aid of a keyboard (System TR 86). The visual information is further processed to topological matrices.

A similar method enabled Ziegler and Boll[20] to input an average of 30 compounds per hour. The advantage of the projection-screen process is that incorrect drawing of the formula can usually be rectified at the stage of projection. These latter methods enjoy the further advantage of keyboard input, which with some practice can be operated more quickly than drawing onto a tablet.

The second method, using a code, appears to be faster. The speed of input for WLN varies from 30,[21] 50,[18,22] up to 150[23] compounds per hour; for the IUPAC notation, up to 100 compounds per hour.[6] There is no information concerning the type of compounds coded.

The input method of Gasteiger and Jochum[24] is more concise than topological bonding tables in general but, nevertheless, describes every single atom. Few rules are called for. However, for general use these rows of symbols are too obtuse and too voluminous compared to other linear notations.

COMPUTABLE CODING OF COMPOUNDS ON THE BASIS OF THE BEILSTEIN SYSTEM

Beilstein Sorting Today. An abstract slip is made for each compound found in an original publication giving name, structural formula, molecular formula, preparation, physical data, and chemical behavior. A first rough sorting, e.g., into the three main categories acyclic, isocyclic, and heterocyclic, is made on scanning the structural formula from the abstract slip. The Beilstein System number, obtained from the index of system numbers, is then noted on the slip. If necessary, a modifying term is added, e.g., in the case of rings, condensed (c) or not condensed (nc).

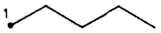
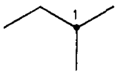
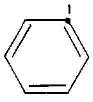
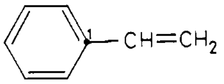
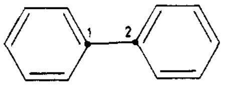
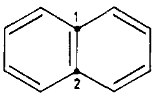
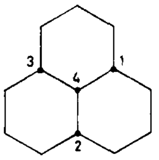
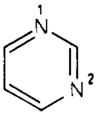
The abstract slips are sorted according to their system numbers by auxiliary staff and then a further precise classification is carried out by specialists. According to the degree of difficulty, from 50 to more than 150 compounds per hour can be classified in this manner. This method is superior to all others from the point of view of its speed, especially as it already includes the sorting process. The deciding factor is now whether it is possible to write down this process in a manner suitable for computerization.

The compounds classified thus can be brought into a mechanically readable form by addition of the CA Registry Number. However, the cost (according to Heller et al.,[25] approximately $3/compound) appears too high. During the 50s a Beilstein fragment code[21a] was developed by the Hoffmann-La Roche Company. A selectivity of 90% was achieved by using 56 descriptors on a punch card; i.e., an average of 10% of cards retrieved were unwanted in the search. Of these descriptors, 17 dealt with the basic structure of the molecule, 5 with the number of C atoms, and the remainder with the functional groups. The speed of coding was similar to that of the WLN, i.e., approximately 100 compounds in 3 h, including checking for mistakes.[21] As compounds could not be described uniquely by this code, a new, systematic method of coding will be demonstrated below.

Table I. Types of Vertices

V1	carbon chain
V2	junction of carbon chains
V3	heteroatom or group of heteroatoms in open chains
V4	carbon ring
V5	junction of a ring and a chain
V6	ring to ring junctions
V7	heteroatom in a ring

Chart I. Fragment Connection Tables

	V type	V no.	at- om	LL[a]	C no.	BV's[b]
	1	1	C	0	4	1
	2	1	C	0	2,1,1	1
$C_2H_5-\overset{1}{O}-C_2H_5$	3	1	O	0	2,2	1
	4	1	C	0	5	4
	5	1	C	0	5,2	4,2
	5	1	C	0	5	4
	5	2	C	1	0,5	1,4
	6	1	C			
	6	2	C	1,1,1	4,4,0	4,4,4
	6	1	C			
	6	2	C	1	3	1
	6	3	C	1,2	3,3	1,1
	6	4	C	1,2,3	0,0,0	1,1,1
	7	1	N			
	7	2	N	1,1	1,3	4,4

[a] LL = lower numbered ligands. [b] BV = bonding value.

I. ANALYSIS OF THE MOLECULE AS A WHOLE

Several workers[26] have employed graph theory, based on the structural formula, to represent and manipulate chemical compounds. The atoms were taken to be the vertices and the bonds the edges of a graph. The H atoms were not usually taken into account (reduced graph). The handling of molecules represented in this way is necessarily expensive and calculation-intensive, and cannot be considered for large amounts of data. For this reason a new, shortened form of graph representation is suggested here, based on the Beilstein formalism. According to this formalism, a compound is fragmented until a precise location in the Handbook is possible. Fragmentation points are taken now as vertices (=Fragment Vertices) of a graph and are labeled with their atom values. The connections between Fragment Vertices are taken as edges of the graph and are labeled with the number of the intermediary carbon atoms. C atoms, C chains, and C rings standing alone are considered to be point graphs. Types of vertices present in molecules are listed in Table I.

To complete the description of a molecule, the bonding values (Table II) and the displacement of the bond from the vertex of origin must be shown.

Table II. Bonding Values

B1	single bond	B5	noncovalent bond
B2	double bond	B6	positive charge
B3	triple bond	B7	negative charge
B4	delocalized bond		

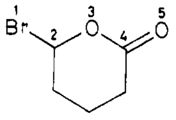
Fragment connection tables (FCT) for several types of vertices are shown in Chart I. Whereas conventional connection tables need one line for each atom, FCT's often need only one line for the whole molecule.

A canonical numbering of the fragment vertices is necessary to render FCT unambiguous. The lowest number is assigned to that vertex with the

1. highest vertex value
2. highest atom value (or function value)
3. highest atom values of ligand vertices
4. highest C number labeled to the vertex
5. highest C number labeled to the ligand vertices
6. highest neighboring vertex number

The following examples may serve for illustration: Example 1:

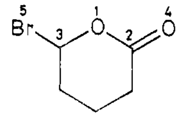
Arbitrary numbering of vertices:



no.	V type	(1) order	atom	(2) order	atom	(3) order
1	3	3	Br	4		5
2	5	2	C	2	Br	3
3	7	1	O	1		1
4	5	2	C	2	O	2
5	3	3	O	3		4

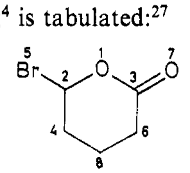
Consideration according to: (1) the V type produces partial ordering into three classes; (2) the atom value produces a further ordering into four classes; (3) the atom values of the adjacent ligands produces the final ordering into five classes.

Resulting canonical numbering of vertices:



	FCT			
no.	atom	LL	C no.	BV's
1	O			
2	C	1	0	1
3	C	1,2	0,3	1,1
4	O	2	0	2
5	Br	3	0	1

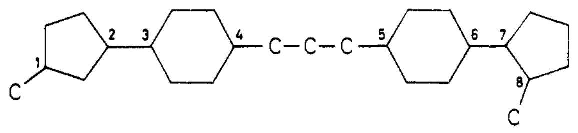
For comparison, a conventional connection table using the Morgan numbering[4] is tabulated:[27]



atom no.	atom	ligand	bond
1	O		
2	C	1	1
3	C	1	1
4	C	2	1
5	Br	2	1
6	C	3	1
7	O	3	2
8	C	4	1
ring closure 6-8			1

Example 2:

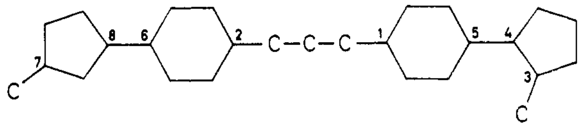
Arbitrary numbering of vertices:



no.	V type	atom	C no.	(3) order	C no.	(4) order	C no.	(5) order
1	5	C	2,1,1	5		6		7
2	5	C	2,1,0	6		7		8
3	5	C	2,2,0	4	2,1,0	5		6
4	5	C	3,2,2	1	2,2,0	1	2,1,0	2
5	5	C	3,2,2	1	2,2,0	1	3,0,0	1
6	5	C	2,2,0	4	3,0,0	4		5
7	5	C	3,0,0	3		3		4
8	5	C	3,1,0	2		2		3

Consideration according to: (1) the V type produces no ordering, (2) the atom value produces no ordering, (3) labeled C numbers produces partial ordering into six classes, (4) labeled C numbers of the adjacent vertex ligands produces a further ordering into seven classes, (5) labeled C numbers of the next adjacent ligands produces the final ordering into eight classes.

Resulting canonical numbering of vertices:



	FCT		
no.	atom	LL	C no.
1	C		
2	C	1	3
3	C	0	1
4	C	3,3	3,0
5	C	1,1,4	2,2,0
6	C	2,2	2,2
7	C	0	1
8	C	6,7,7	0,2,1

II. ASCERTAINMENT OF THE MAIN FRAGMENT

The main fragment is that fragment with the lowest classification value (see Table III) which results on heteroscission V3 throughout the whole molecule. The main fragment always retains the heteroatom following V3 fission.

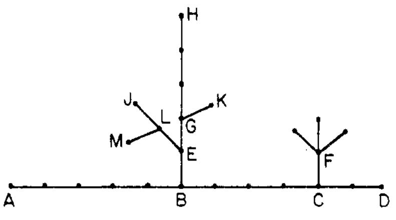
Definitions:

Ring Layer: A fused ring system is regarded as a layer lattice. A linear chain of ortho-fused rings forms a ring layer. The ring chain is defined as linear when the next ring is situated on the respective outermost opposite ring edge.

Main Layer: That ring layer that (a) contains the most rings and (b) least divides the molecule.

Further definitions are shown in Diagram I (cf. ref 28).

Diagram I

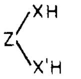
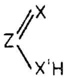
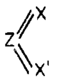
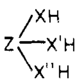
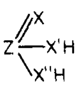


B = root of tree 1, C = root of tree 2
A-D = main chain
B-H = main branch (1st order)
E-J, G-K = branches of 2nd order, L-M = branch of 3rd order
B, C, E, F, G, L = junctions

Table III. Classification Value

no.	V type		position
1	7	heterocyclic [atomic value (Table IV), number of heteroatoms]	0
2	4	isocyclic	0
3	3	functional groups [function value (Table V), number of functional groups]	3
4	B	number of unsaturations	8
5	1+4+7	C number	10
6	4+7	number of rings	12
7	7	number of heterocyclic rings	14
8	6	number of ring-to-ring junctions	15
9	4	size of ring (descending)	17
10	7	size of heteroring (descending)	25
11	4+7	further ordering of fused ring systems:	
		(a) number of rings on main layer	31
		(b) largest number of rings not divided by the main layer	32
		(c) number of layers parallel to main layer	33
12	7	further ordering of heterocyclic ring systems: arrangement of heteroatoms	
13	5	length of C chains lying between or on the rings	
14	1+2	further ordering of open-chained branches:	
		(a) length of main chain	
		(b) number of junctions	
		(c) arrangement of junctions	
15	B	arrangement of multiple bonds	
16	3	number and arrangement of different types of functional groups	
17	3	number and arrangement of nonfunctional inorganic substituents	
18		classification value and arrangement of organic substituents	

Chart II

	type	symbol	
1.	Z—XH		e.g., -OH
2.	Z═X	X	(*ox*o type)
3.		D	(*di*), e.g., acetals
4.		C	(*c*arboxylic acid type)
5.		Q	(*q*uadruple)
6.		T	(*t*riple), e.g., orthoesters
7.		Z	(**Z**), e.g., phosphates
8.	Z≡X	L	(nitri*l*e)

Inorganic groups are specified according to the general types in Chart II (Z = central atom). Combining the symbols in this order leads to other types. Symbols of central atoms other than C are given before the type symbol. Free hydrogens transcending the lowest valence level of the central atom are given as H's behind the atomic symbol. Symbols of X atoms other than O are given behind the type symbol in order of descending bond values (see Table IV). This representation of inorganic groups allows calculation of valence and oxidation number of the central atom by computer.

Other functional groups, as listed in "How to Use Beilstein"[29a] or in the "Beilstein Guide",[29b] and nonfunctional groups are similarly symbolized.

Table IV. Important Functional Groups

	symbol	function value
$-OH$	A (*a*lcohol)	1
$-SH$	S	1,1
$=O$	X (*ox*o)	2
$=S$	XS	2,1
$=NH$	M (*im*ido)	2,2
$=O(OH)$	C (*c*arboxylic acid)	3
$=O(NH_2)$	CN	3,4
$=N(NH_2)$	CMN	3,5
$\equiv N$	L (nitri*l*e)	3,6
$=S(O)(O)(OH)$	SXC (sulfonic acid)	4,2
$-NH_2$	N	5
$-NHNH_2$	NN	7,1
$-N=NH$	NM	7,2
$-PH_2$	P (phosphines)	8,1
$-PH_4$	PHH (phosphoranes)	8,4
$-P(O)(OH)(OH)$	PZ (phosphonic acids)	8,9

Table V. Atomic and Pseudohalogenic Values[a]

1,1	F		2,4	Te	4,4	Pb
1,2	Cl		3,1	N	5,1	B
1,3	Br		3,2	P	5,2	Al
1,4	I	Code:	3,3	As	6,1	Be
1,5	(NO)	NX	3,4	Sb	8,1	Cu
1,6	(NO_2)	NQ	3,5	Bi	9,1	Zn
1,7	(N_3)	NMM	4,1	Si	9,3	Hg
2,1	O		4,2	Ge	13,1	Cr
2,2	S		4,3	Zn	15,9	Pt
2,3	Se					

[a] First number (1-7), 7th-1st main group; (8-15), 1st-8th subgroup of the periodic system.

The Beilstein sequence is given by the atomic values of the H or OH replacing groups. Table V (cf. ref 29b) may serve also for producing the Beilstein sequence of heterocyclic compounds.

III. RULES FOR ENCODING (EXAMPLES: CHART III)

1. Fragmentation. For fragment at heteroatoms of type V3, find the main fragment according to section II.

2. Base Fragments. Further fragmentation of the main fragment yields the base fragments: a ring or ring system and the attached heteroatoms, or a C chain with attached heteroatoms. The code of the base fragment is the C number and the symbols of functional groups or other heteroatoms.

The existing international atomic symbols are used except for Cl and Br; for these, WLN symbols are used: G and E. Two-character atomic symbols and numerals are preceded by the symbol ":".

3. Description of Cyclic/Noncyclic State. Cyclic and acyclic parts of the molecule are separated by a dash. Components after an even number of dashes (including zero) are cyclic, after an odd number acyclic (examples 7, 9, 10, etc.).

4. Numbering. The numbering follows the general rules of chemical nomenclature, in particular, "The Ring Index".[30] Locants appear in parentheses after the corresponding code symbol. The number of consecutive locants or the numeral behind the multiplication symbol ! (example 18) denotes the degree of occurrence of the preceding symbol. Locant 1 can be omitted where there is only one code symbol so indicated (example 3). Ambiguities in the second terminus of an unsaturated bond are removed by specification of the second locant after the sign "." (= end of primary locants).

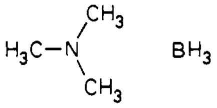
5. Fused Rings. The codes for each single ring of a ring layer are in immediate consecutive order (examples 11-15). A new ring layer is preceded by the symbol &. When a new ring is attached to two or more rings of a system, the lowest number of the previous layer (numbered from the right) is put

Chart III

H_2SO_4

1

−SCC



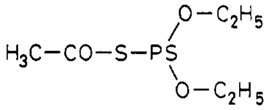
2

(a) −1N/11.B
(b) −N/1(!3).B

$H_3C-[CH_2]_{12}-OH$

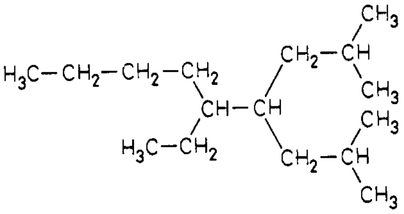
3

−:13A



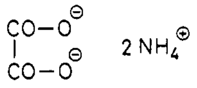
4

−2COSPZS/2(ZZ)



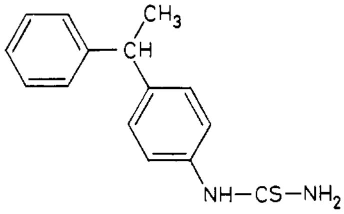
5

−93(4)1(2′)2(5)1(2)



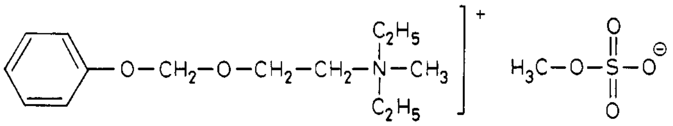
6

−2C(−)(12).N(+)(!2)



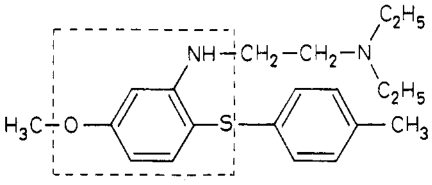
7

6–11–6–N(4)/1CSN



8

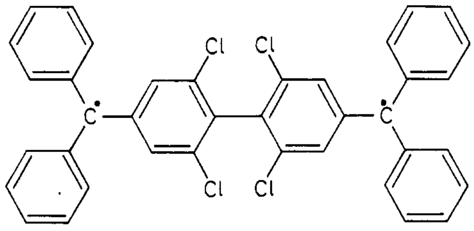
6–A/1A//2N(2+)///2(NN)1(N).1ASXC(−)



9

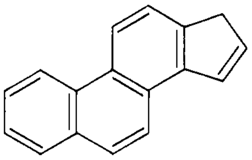
(dotted lines = main fragment)

(a) 6–A(1)N(3)S(4)/2(N)N(2)//22/–6(S)–1(4)/1(A)
(b) 6–A(1)N(3)S(4)/–6(S)–1(4)/2(N)N(2)//22
(c) 6–A(1)[1]N(3)[2N(2)/22]S(4)[–6–1(4)]



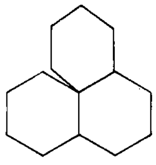
10

(a) 6–1(*)–6(!2)–G(26)–6(1)–G(2′6′)1(4*)–6(!2)
(b) −1(*)6(!3)–G(26)–6(1)–G(2′6′)–1(4′*)–6(!2)



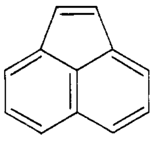
11

66&65H(:17)



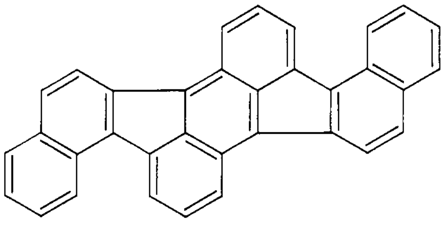
12

66&6(1R)



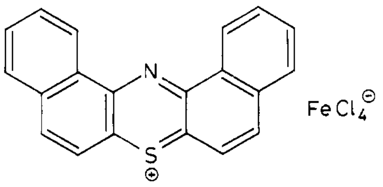
13

66&5(1.2)



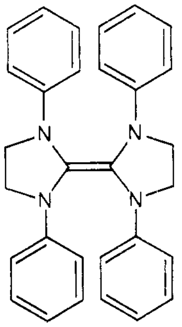
14

65656&6(2.3)6&&6(3.4)6(5)



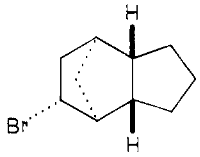
15

64N(:14)S(7+)6&6(13).
:FEG(!4)(−)



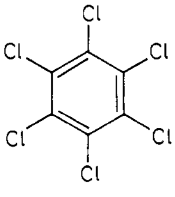
16

3N(13)H−=(2.2)–3N
(13)H/6(131′3′)



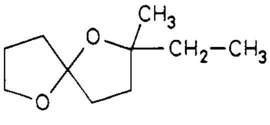
17

5[1](47.F)55H–E(5.C).
H(3A7A.TT)



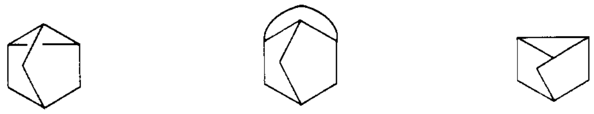
18

6–G(!6)



19

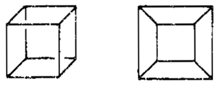
4O+(2.2)4OH–2(5)1(5)



20

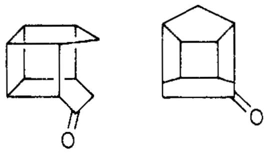
5[1](14)5[1](26)3H

Chart III (*Continued*)



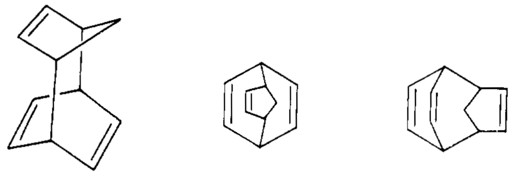
21

444&4(1.23)&&4(1.23)H



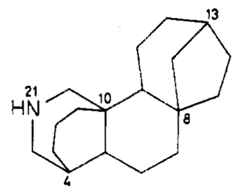
22[31]

645&4(1.23)&&4(1.23)H–X



23[31]

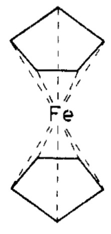
6[2](14)7[1](7:10)5H=(258)



24

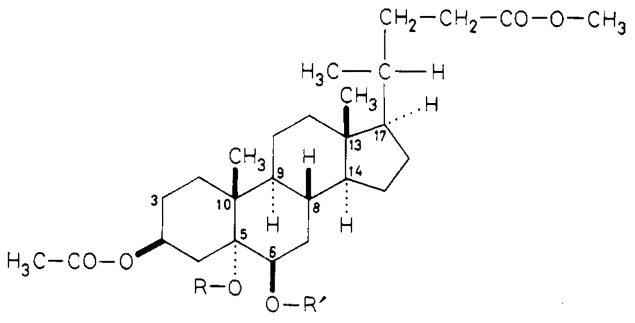
veatchane[32]

5N:21)[1](4:10)66&6[1](8:13)5H

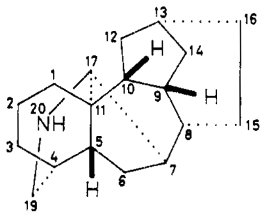


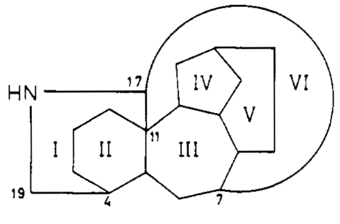
25

5.–:FE.–5



26a, R = H, R′ = CH_3

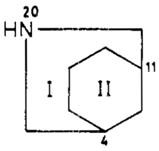




27

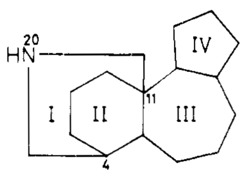
28

5N(:20)[1](4:11)67&5[2](8:13)6[7](7:17)9H



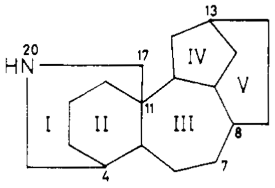
29

5N(:20)[1](4:11)6H



30

5N(:20)[1](4:11)67&5H



31

5N(:20)[1](4:11)67&5[2](8:13)6H

in parentheses behind the ring code. Additional numbering is given behind "." (examples 13, 14, 21, and 22). Edges are numbered R, S (*s*tandard, to be omitted), T, etc. or L, M, N, etc., starting from the nearest *r*ight (upper) or nearest *l*eft (*l*ower) node (example 12). The molecule is placed in such a way that the main layer is situated at the bottom, with most rings in the upper right quadrant (see "Ring Index"). Coding begins with the main layer and moves first upwards through the various ring layers and then, separated by a double &, downwards (example 14). Bridging atoms between rings are shown within square brackets followed by the locants of the bridge termini (example 17). Three-dimensional structures are converted to a two-dimensional form (examples 20–23).

6. Multiple Bonds. Cyclic parts are regarded as principally unsaturated, acyclic parts as principally saturated. Deviations are marked by the symbols H (saturation point), = (double bond), and == (triple bond); these can be combined in any way. When "H" is not followed by a locant, then everything preceding this symbol is regarded as saturated (examples 12, 17, 19, 24).

7. Branched C Chains. C-Chain trees branching off the main chain are coded successively in descending size of the main branch (see Diagram I: first tree 1, then tree 2). Codes of the branches of a tree are given in ascending order (see Diagram I). The order is shown by the corresponding number of quotation marks after a locant (examples 5, 7, and tree 1 of diagram I: $C_{BH}C_{EJ}(1')C_{GK}(2')C_{LM}(1'')$).

8. Further Heterofragments. After encoding the main fragment according to rules 2 to 7, the remaining fragments are coded: fragments of one fragment tree in ascending order, followed by the fragments of the next fragment tree. The fragment with the lower resulting classification value always retains the heteroatom following V3 fission (example 9). The order, defined as the number of heterovertices between the fragment and the main fragment, is shown in the corresponding number of oblique strokes preceding the fragments of a new order level or a new fragment tree. If necessary, it is shown at which functional group the fragment is bonded (examples 4, 8, and 9). Fragments or fragment trees may also be shown in square brackets immediately after the functional group concerned (example 9c). This operation may be performed only once per functional group.

9. Salts, Radicals, Molecular Compounds. Localized charges (signs + and −) and positions of radicals (sign *) are shown immediately after the relevant locant. Nonlocalized charges are shown either at the end of the fragment concerned

Table VI. Stereo Symbols

α, β (on steroids)	A, B
cis, trans, *re*ference atom	C, T, F
endo, exo	D, X
syn, anti	N, I
E, *Z*	E, Z (after locant for double bond)
R, *S* (according to CIP rules)	R, S

or at the end of the molecule. Parts of molecular compounds, charged molecule parts, and noncovalent bonded molecule parts are separated by the sign "." (examples 6, 8, 15, and 25).

10. Oligomers and Polymers. The position of an open bond in the molecule is marked by * after the corresponding locant. The degree of polymerization is given behind the symbol "." (=end of description of constitution) in parentheses behind the multiplication symbol "!".

11. Stereochemical Conventions. Stereo symbols from Table VI are shown within parentheses after the locants and after the sign "." (=end of primary locants) (example 17). Descriptions which apply to the whole of the molecule are shown in plain text after the sign "." (=end of description of constitution).

12. Sequence of Symbols and Fragments. The fragments are handled in the same order as they are connected in the molecule. For computerization canonical encoding is not necessary. The computer program puts each structure descriptor in its correct place. If a canonical code is desired, symbols and fragments must be arranged in order of their tabulated values and vertex values, respectively. The rule of setting substituents in square brackets behind a functional group becomes invalid.

Symbols Used Throughout.

(....)	locants
:	two-character atom or number
/	fragmentation point according to V3 (splitting at heteroatoms)
-	change of cyclization state—even number: cyclic, odd number: acyclic even number: cyclic, odd number: acyclic
&	start of a new ring layer
+	spiro junction
[...]	description of bridge atoms
[...]	behind a functional group: fragment, fragment tree, heteroatom, or heteroatom group bonded to this functional group
=	double bond
==	triple bond
(+,−)	positive or negative charge
(*)	radical or open bond
.	end of description of constitution of an entity of covalent bonded atoms
(.)	end of primary locants
(!...)	multiplication sign

IV. EXAMPLES FOR ENCODING

(a) Compound **26a**:[33] Ascertain the heterovertices V3 and examine the resulting fragments A–D:

A: the ring skeleton with acyclic side chains
B: CH_3CO
C: CH_3
D: CH_3

Fragment A is isocyclic and is given the lowest classification value according to Table III. It follows that A is the main fragment.

The heterovertices are considered together with the main fragment, which now has the following functional groups: three OH groups and one O(OH) group. The main fragment contains the rings 6665. As we are dealing with a fused system with two ring layers, the code for the basic skeleton is 66&65.

An H follows as there are no double bonds, and then a dash after which the acyclic components of the main fragment are shown. The functional groups of the ring system are stated and numbered according to "The Ring Index". The code symbol for the OH group is A (Table IV):

66&65H-A(356)

Now follow the acyclic C chains with the remaining functional groups:

66&65H-A(356)4(:17)C(4)1(1′)1(:10:13)

This then is the description of the main fragment (fragment A).

The remaining fragments are all of the order 1. Methyl groups (fragments C and D) are attached to the acid function and the hydroxy function at position 6:

66&65H-A(356)4(:17)C(4)1(1′)1(:10:13)/1(AC.6)

The acetyl group (fragment B) is attached to the hydroxy function at position 3:

66&
65H-A(356)4(:17)C(4)1(1′)1(:10:13)/1(AC.6)/2(A.3)X

Thus the constitution of the molecule is fully described. In order to describe the configuration we must add the stereosymbols A and B (see Table VI):

66&
65H-A(356.BAB)4(:17)C(4)1(1′)1(:10:13.BB)/1(AC.6)/
2(A.3)X.H(89:14:17.BAAA)

These stereocodes may be listed under special abbreviations in the steroid class.

(b) Take as a second example compound **27**, the Aconitan skeleton.[34] For clarity the compound is represented in such a way that the bonds do not overlap. This can be done with the aid of curved bond lines (**28**). Encoding begins with the heteroring I and the attached ring II (**29**). A bridge atom is situated between bridge termini 4 and 11. Rings III and IV follow, whereby ring IV starts a new ring layer (**30**). Two new bridge atoms between termini 8 and 13 create ring V (**31**). Seven new bridge atoms between termini 7 and 17 give rise to ring VI. Thus the constitution of the compound is fully described.

V. TIME EXPENDED ON CODING

A notation to be used in practice must be deducible simply and quickly from the structural formula. A suggested new notation should offer appropriate advantages over existing notations. This will now be examined.

(a) Theoretical Aspects. We proceed from the supposition that the time required for contemplation of a structural formula and its conversion to a linear form is the same in each case. Each atom and each structural characteristic must be considered with each notation. The time added to this "fixed time" can, however, vary widely. This can be reduced when (1) connected parts of the molecule are coded in conjunction; (2) the code is based on customary rules of nomenclature; (3) symbols familiar to the chemist are employed; (4) as few rules as possible are to be observed; (5) these rules conform to the common chemical behavior of molecules; (6) the rules are easy to understand; (7) special signs used are easy to memorize; (8) the use of special signs is kept to a minimum; (9) the code is flexible, i.e., able to adjust to the respective structural particulars; (10) the rules accommodate all possibilities; (11) meaningful abbreviations for derivatives are possible (substituent R); (12) frequently occurring compounds and basic skeletons are allocated as simple a code as possible.

An attempt was made to deal with all aspects of the above items, paying special attention to items 2, 3, 4, 6, 9, 11, and 12. Note especially the easy handling of the most frequently occurring cyclic compounds. There is no other notation which uses only 6 for benzene, 66 for naphthalene, and 5N for pyridine. Where other advantages over previous notations lie must be examined from case to case; each notation offers certain advantages and meaningful comparison has up to now not been possible. There is a shortage of information about the types of compounds encoded. As suggested by Neubert,[35] a parameter is here used to remedy this situation, namely, the number of nonhydrogen atoms (NHA). This number should be an approximation for the difficulty of coding a compound.

Several compound files of Beilstein Handbook were analyzed by computer in order to reach an average value of NHA. The following values resulted:

4638 acyclic compounds:	16
3053 isocyclic compounds:	17
4556 heterocyclic compounds:	20

Consequently an average value of NHA = 18 was obtained from approximately 12 000 compounds.

(b) Experiments

Experiment 1: Materials used: 222 abstract slips from current output. Type of compounds: mixed, but predominantly monocyclic and heterocyclic compounds with several functional groups. Execution: Coding is carried out by a chemotechnician. Training: 14 hours. Coding time: 55 compounds per hour (including writing down the molecular formula). NHA number: 16.

Experiment 2: Materials used: 102 mixed abstract slips from current output. The number of heterocyclic and fused-ring compounds was proportionally higher than for experiment 1. Execution: author. Coding time: 65 compounds per hour.

Experiment 3: The material from experiment 2 was arranged according to the abstract slip numbers. This is the order in which the abstracts are written and delivered by the abstractors. Execution: author. Coding time: 84 compounds per hour.

Experiment 4: Materials used: 34 abstract slips sorted according to the Beilstein System. The compounds in question were piperazine derivatives with varying organic substituents on both N atoms. Execution: author. Coding time: 92 compounds per hour (including writing down the molecular formula).

Experiment 5: Materials used: 63 compounds taken from different chapters of the WLN Handbook.[7] Execution: author. Coding time: 70 compounds per hour.

Experiment 6: The material from experiment 2 was "systematized", i.e., was given the Beilstein system number and modifying term (see above). Execution: a practiced chemotechnician. Systematization time: 204 compounds per hour. Total time including subsequent precision-sorting: 160 compounds per hour.

Errors:

(a) systematic errors at the coding stage

(b) errors due to carelessness

(c) errors at the stage of computer input

Total error quota of experiment 1: 15% (prior to correction). A common source of mistakes on typing was the inclusion of customary commas between locants. For this reason we have developed a method for the representation of locants without commas (see coding rule 4). All errors could be detected by computer program.

Evaluation:

The material represented a cross-section of the multitude of organic chemical compounds. The number of compounds coded per hour could be increased when the material was first arranged in order. Accordingly, in the case of derivatives only the variation in substituent required coding (using "R" for the unaltered stem). Compared to the Beilstein systematization method (experiment 6), the expenditure of time was twice as high, caused partly by having to write down the code.

Further, unlisted experiments showed an average time of encoding between 50 and 70 compounds per hour under normal production conditions (including interruptions of work). How far coding time could be reduced by better training of staff was not investigated. The error quota should be lowered significantly with more practice.

Table VII. Locant Sequence Code

rings in fused ring systems	1
junctions of fused rings	2
branchings	3
heteroatoms in rings	4
double and triple bonds	5
functional groups	6
organic substituents	7
inorganic coupling components[29]	8
halogens, pseudohalogens, S, Se, Te	9

REPRESENTATION IN THE COMPUTER

(a) BUS Number. A computer program (see below) converts the code to the BUS number (BUS = *B*eilstein *U*nique *S*equence), a number unique and unambiguous for each chemical compound. Each structural characteristic is allocated a certain position, which depends on the Beilstein order of precedence (Table III). Table III also shows the position in the computer representation (last column). Classification values 1–11 are thus allocated a total of 34 spaces.

Locants are a further feature of classification in the Beilstein System. Thus there follows in positions 34–49 a locant sequence code (Table VII) characterizing the following locant fields of each 10 spaces. Space 1 of a locant field is the order field; spaces 2–4 are the identification field; spaces 5–10 are the locant field proper. The identification field shows, for instance, for code = 3 the length of the branching chains, for code = 6 the number of functional groups, for code = 9 the halogen/chalcogen code (F = 0, Cl = 1, Br = 2, I = 3, NO = 4, NO_2 = 5, N_3 = 6, S = 7, Se = 8, Te = 9) in correct sequence. Two spaces are allocated to each locant: the first for alphanumeric representation (allowing 36 positions in practice), the second for a stereo symbol code. Thus three locants can be accommodated in a locant field. If this is insufficient, a further locant field is defined.

Derivative fields, defined by a 7, consisted of 20 spaces in practice. Space 1 shows the order of the derivative $(10 - x)$. A second derivative field may be defined if 20 spaces are inadequate to the description of a heterofragment.

The BUS number for compound **26a** is described in detail below, but without the stereo symbol code for better clarity. Compound **26a** is followed by **26b** and **26c** in the Beilstein Handbook.[33] The altered lines of the BUS number are shown in Chart IV.

(b) Fragment Jump Numbers. The following process may be used for structuring a larger file which can be searched on the basis of partial structures. Each fragment is stored in sequence according to its BUS number, and is issued with a key number (ISAM number) by computer. The substituents appropriate to a main fragment are given as jump numbers in that field originally reserved for description of organic derivatives, namely, in locant field 7. When no organic substituents exist, the key number of the main fragment (=single fragment) may be repeated.

In the case of retrieval of substructures, the SNN of the sought-for structure element is input, which is converted via

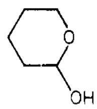
Chart IV. BUS Numbers of **26**[a]

26a	0		2		4		6		8		10		12		14		16		18		line
0			0			1	2	9		5	2	4		4	0		6	6	2	7	1
20	0											2	2	1	3	3	6	6	7		2
40											0	4	1	1	Q		J		M		3
60	1	1			A						1	1			D						4
80	0	3			C		E		F		0	C	A	A	D		C		F		5
100	9												1								6
120	9							1	7		1		2								7
140	9												1								8
26b																					
120	9							1	7		1		2								7
140	9							1	7		1		2								8
26c																					
20	0											2	2	1	3	3	6	6	7	7	2
100	0	A			E																6
120	9												1								7
140	9							1	7		1		2								8
160	9							1	7		1		2								9
180	9												1								10

[a] a, R = H, R′ = CH_3; b, R = H, R′ = $COCH_3$; c, R = CH_3, R′ = $COCH_3$.

BUSEN (see below) to the corresponding BUS number. All fragments which contain this structural feature are found by checking the appropriate columns. Their ISAM numbers point to the molecules containing this fragment. An obvious advantage of this type of file-structuring is the conservation of storage space, since fragments which occur repeatedly need not be stored in a redundant manner. New molecules or fragments to be added to the file may cause automatic alteration of key numbers and fragment jump numbers.

(c) Tautomeric Compounds. A unique representation of tautomeric compounds may be achieved by allocating these to a defined place according to special rules. Coding one or the other form should result in the same place in computer representation. Take, for example, the oxo–cyclo tautomerism of the following compound:



initial state I

$HO-[CH_2]_4-CHO$

final state F
(= Beilstein registry place)

BUS numbers (places 1-12):

I		1					1			1		5
F		0				4	1			1		5
		RC				FC				U	CN	

Detecting a hydroxy group neighbored to ring-O, the computer program jumps to a tautomerism subroutine and gives rise to following changes in representation:

Ring code: 1 → 0

Functional group counter: + one oxo group.

These changes formally describe a chemical reaction (ring opening or closing). The algorithm for describing this type of reaction is as follows:

$$I \rightarrow F\text{: (a) } RC - 1 \rightarrow RC$$

$$\text{(b) } FC + FC(=O) \rightarrow FC$$

$$F \rightarrow I\text{: (a) } RC + 1 \rightarrow RC$$

$$\text{(b) } FC - FC(=O) \rightarrow FC$$

(RC = ring type code, FC = functional group code, U = unsaturations, CN = number of carbon atoms). Other tautomeric changes may be handled analogously.

Computer Program. (a) Beilstein Sequence. The program BUSEN (= *B*eilstein *U*nique *S*equence *N*umber), written in Assembler, ran on a Siemens 4004/BS 2000 computer. The main routine separates the read-in code into cyclic and acyclic components. Each dash causes a jump to the respective subroutine. Reading the signs triggers counters in specifically allocated fields, while signs such as [and (cause a jump to the appropriate subroutine. Simultaneously occupied bit fields can be used later in different ways.

Functional groups are recognized only in the acyclen subroutine. They are added up according to their Beilstein precedence of order. A single letter in the cyclen subroutine, on the other hand, precipitates an error message (apart from the single figured atom symbols). Such syntactic error recognitions are built into the program in several places.

A molecular formula counter operates simultaneously with sign recognition. For example, if an N is recognized, then the N-counter increases by 1. At the end of the run the molecular formula arrived at in this way is compared with the input molecular formula. Mistakes arising from false locants cannot be recognized in this manner. A check digit may be calculated on input, and this may be examined by the computer program. Practically all mistakes are eliminated by checking for (a) syntactic errors, (b) correct molecular formula, and (c) correct check digit.

The calculated fragment BUS numbers are compared with each other to establish the correct sequence. The compound BUS numbers (CPU-time ca. 0.01 s per compound) are then listed and sorted in ascending order. The code and other data (e.g., the abstract slip numbers) are sorted in a second file (requiring less than 100 bytes per compound) with the aid of the simultaneously sorted key numbers. Output is a list of sorted codes and the corresponding abstract slip numbers (or, e.g., the microfilm addresses). In this way the aim of the Beilstein precision classification per computer is achieved.

(b) Connection Table of Atoms. A subprogram of BUSEN is the program dealing with the generation of atom connection tables. Here the numbering according to the Ring Index has also been used. Special numbering for trivial names is, however, not permitted. The program so far embraces only simple cyclic systems (including fused rings). It may not always be possible to adhere to the Ring Index numbering in the case of more complicated ring systems.

For producing an unequivocal connection table it is necessary to generate an unequivocal numbering, e.g., the Morgan algorithm,[4] the CANON algorithm of Schubert and Ugi,[36] the Nodal Nomenclature of Lozac'h et al.,[37] an atom numbering derived from the fragment vertex numbering proposed in this work, or an algorithm considering the environments of a focus-atom (e.g., the ELCO module in the DARC system of Dubois[38] and the concept of path graphs of Randić and Wilkins[39]). First atttempts were encouraging.

Possibilities for Direct Access. It is frequently necessary to extract a certain compound from storage for checking purposes during the Beilstein manuscript processing. Use of the SNN offers only limited possibilities, as the code for one and the same compound can vary. This would result in the same problems as with nomenclature. Furthermore, a great deal of calculation time would be required for the computer to search sequentially for a certain compound through a file of several million entries.

The following methods show potential and have in some cases been tested on existing small files:

(a) On calculation of the BUS number in the computer the SNN is simultaneously canonicalized and stored in its new form; i.e., the computer checks the sequence of the code symbols and fragments and rearranges their symbols if necessary. The canonicalized code can then be recalled direct from the master file. It is, however, required first to canonicalize the asking-code with the aid of the same computer program. Moreover, the size of the file necessitates screening, e.g., by checking the bit fields or the molecular formulas.

(b) The master file, consisting of nonunique SNN, molecular formulas, and bit fields, is screened as under (a). The output codes are converted to the BUS number and compared with the BUS number of the asking code.

(c) Not the code but the BUS number is stored, and these are searched. The strict ascending order enables particular methods for direct access. However, this form of storage calls for ~200 bytes storage space per compound.

(d) Fragment BUS numbers are stored and connected with jump numbers. Checking certain columns of the master file serves as screening for retrieval.

Conclusion. The information processing system for chemical compounds here suggested is based on the principles of common nomenclature, and therefore does not represent a fundamentally new system, unlike most other notations described hitherto. There are opportunities for the conversion between graphic, SNN, nomenclature, fragment, and connectivity representations.

The symbols used throughout are easy to retain and follow a strong partitioning: letter symbols are reserved for atom symbols and functional group symbols, numeric signs are used for C numbers, and special signs are used for characterizing bonds and various junctions. Within square brackets, these symbols describe ring bridges. Connections, multitudes, charges, and configurations are described with parentheses, using generally well-known symbols. This clear partitioning helps even untrained chemists to handle SNN without major difficulty.

The user is furthermore not confronted with the awkward decision concerning the coding of tautomeric forms. The possibility of tautomerism is checked by the program BUSEN, and the compound is subsequently assigned its unique and proper place. Similarly, the computer can perform the task of seeking the main fragment. The possibility that the SNN could be rendered unique (in the sense that for any one structure, only one SNN could apply) by means of the sequences listed in the tables was not investigated, since this is neither central to, nor necessary for, the purposes of Beilstein. The computer program outlined achieves its purpose and enables a classification of great precision.

The concept of fragment vertices, as derived from the Beilstein System, allows connection tables with significantly lower storage requirements than hitherto. It shows perspectives for a simple mathematical representation of molecules and will allow further applications.

It is not possible in this paper to deal with all classes of compounds to be encoded. The construction of SNN allows, however, for the coding of other compound classes not dealt with here. This and further experiences in handling SNN may be demonstrated in further publications.

ACKNOWLEDGMENT

I am grateful to Professor Dr. R. Luckenbach for rendering this work possible, E. Lovas for encoding the compounds and helpful discussions, and Dr. A. J. Lawson for translating this work and valuable criticism.

REFERENCES AND NOTES

(1) S. Rössler and A. Kolb, "The GREMAS System, an Integral Part of the IDC System for Chemical Documentation", *J. Chem. Doc.*, **10**, 128 (1970).

(2) W. Steidle, "Möglichkeiten der mechanischen Dokumentation in der organischen Chemie", *Pharm. Ind.*, **19**, 88 (1957).

(3) W. Nübling and W. Steidle, "Der Dokumentationsring der chemisch-pharmazeutischen Industrie; Ziele und Methoden", *Angew. Chem.*, **82**, 618 (1970).

(4) H. L. Morgan, "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service", *J. Chem. Doc.*, **5**, 107 (1965).

(5) H. Grünewald, "IDC and its Methods of Operation", *Pure Appl. Chem.*, **49**, 1855 (1977).

(6) H. F. Dammers and D. J. Polton, "Use of the IUPAC Notation in Computer Processing of Information on Chemical Structures", *J. Chem. Doc.*, **8**, 150 (1968).

(7) E. G. Smith and P. A. Baker, "The Wiswesser Line-Formula Chemical Notation", 3rd ed., Chemical Information Management, Inc., Cherry Hill, N.J., 1976.

(8) Chi-Hsiung Lin, "SEFLIN-Separate Feature Linear Notation System for Chemical Compounds", *J. Chem. Inf. Comput. Sci.*, **18**, 41 (1978).

(9) D. Lefkovitz, "A Chemical Notation and Code for Computer Manipulation", *J. Chem. Doc.*, **7**, 186 (1967).

(10) W. J. Martin, "A Readable Chemical Notation", *J. Chem. Inf. Comput. Sci.*, **18**, 199 (1978).

(11) R. G. Dromey, "A Simple Tree-Structured Line Formula Notation for Representing Molecular Topology", *J. Chem. Inf. Comput. Sci.*, **18**, 225 (1978).

(12) J. Ash and E. Hyde, "System for Chemical Retrieval", *Pure Appl. Chem.*, **49**, 1845 (1977).

(13) O. Kennard, F. H. Allen, M. D. Brice, T. W. A. Hummelink, W. D. S. Motherwell, J. R. Rogers, and D. G. Watson, "Computer Based Systems for the Retrieval of Data: Crystallography", *Pure Appl. Chem.*, **49**, 1807 (1977).

(14) E. Garfield and M. Sim, "The Index Chemicus Registry System—Past, Present and Future", *Pure Appl. Chem.*, **49**, 1803 (1977).

(15) E. Meyer, "Eine Maschine zur Verschlüsselung chemischer Strukturformeln für die Dokumentation", *Nachr. Dok.*, **13**, 144 (1962).

(16) G. Ohnacker and W. Kalbfleisch, "CCBF-Ein System zur Computerbearbeitung chemischer und biologischer Forschungsergebnisse", *Angew. Chem.*, **82**, 628 (1970).

(17) E. J. Corey and W. T. Wipke, "Computer-Assisted Design of Complex Organic Syntheses", *Science*, **166**, 178 (1969).

(18) D. A. Koniver, W. J. Wiswesser, and E. Usdin, "Wiswesser Line Notation: Simplified Techniques for Converting Chemical Structures to WLN", *Science*, **176**, 1437 (1972).

(19) H. S. Neubert, "Computer Aided Input of Graphic Information by Keyboarding under Visual Control of Display as Applied to Chemical Structures", Proceedings of the 24th Meeting of the AGARD Technical Information Panel, Oslo, 1971.

(20) E. Ziegler and K. Boll, "Computer Input and Graphical Reproduction of Chemical Structures", *Anal. Chim. Acta*, **103**, 237 (1978).

(21) National Academy of Sciences-National Research Council, "Survey of Chemical Notation Systems", Publication 1150 Washington, D.C., 1964; (a) p 247.

(22) C. M. Bowman, F. A. Landee, and M. H. Reslock, "A Chemically Oriented Information Storage and Retrieval System. I. Storage and Verification of Structural Information", *J. Chem. Doc.*, **7**, 43 (1967).

(23) J. K. Horner, "Low-Cost Storage and Retrieval of Organic Structures by Permuted Line Notations: Small Collections", *J. Chem. Doc.*, **7**, 85 (1967).

(24) J. Gasteiger and C. Jochum, "EROS-A Computer Program for Generating Sequences of Reactions", *Top. Curr. Chem.*, **74**, 95 (1978).

(25) S. R. Heller, G. W. A. Milne, and R. J. Feldmann, "Quality Control of Chemical Data Bases", *J. Chem. Inf. Comput. Sci.*, **16**, 232 (1976).

(26) A. T. Balaban, "Chemical Applications of Graph Theory", Academic Press, London, 1976.

(27) R. G. Dromey, "A Linked-Path Connection Table with Substructural Atom-Ordering", *J. Chem. Inf. Comput. Sci.*, **19**, 37 (1979).

(28) E. Cayley, *Ber.*, **8**, 1056 (1875).

(29) (a) Beilstein Institute: "How to use Beilstein", Springer-Verlag, Berlin-New York, 1979; (b) O. Weissbach, "The Beilstein Guide",

Springer-Verlag, Berlin, Heidelberg, New York, 1976.
(30) "The Ring Index", 2nd ed., 1960.
(31) W. Grimme, W. Mauer, and G. Reinhardt, *Angew. Chem.*, **91**, 254 (1979).
(32) Beilstein EIII/IV, **20**, 3329.
(33) Beilstein EIII, **10**, 2161; Ring Index No. 4781.
(34) Beilstein EIII/IV, **21**, 2673.
(35) H. S. Neubert, private communication.
(36) W. Schubert and I. Ugi, "Constitutional Symmetry and Unique Descriptors of Molecules", *J. Am. Chem. Soc.*, **100**, 37 (1978); W. Schubert and I. Ugi, "Darstellung chemischer Strukturen für die computergestützte deduktive Lösung chemischer Probleme", *Chimia*, **33**, 183 (1979).
(37) N. Lozac'h, A. L. Goodson, and W. H. Powell, "Die Nodalnomenklatur-Allgemeine Prinzipien", *Angew. Chem.*, **91**, 951 (1979).
(38) J.-E. Dubois, "Structural Organic Thinking and Computer Assistance in Synthesis and Correlation", *Isr. J. Chem.*, **14**, 17 (1975), and references therein.
(39) M. Randič and C. L. Wilkins, "Graph-Based Fragment Searches in Polycyclic Structures", *J. Chem. Inf. Comput. Sci.*, **19**, 23 (1979).

LETTERS TO THE EDITOR

BIBLIOMETRICS AND THE CLINICAL FATE OF DRUGS

Dear Sir:

Windsor, in his article, "Using Bibliometric Analysis of Patent Literature for Predicting the Clinical Fate of Developing Drugs",[1] claims that there is a relation between the clinical success of a drug and traits in the literature about this drug. In the aforementioned article he concentrates on patent literature, for which he develops some mathematical techniques. In an earlier article,[2] remarkably enough not cited in the present one, he presented traits in the journal literature, which he related to the clinical success of a drug. Regarding the earlier article, I wrote a "Letter to the Editor"[3] in which apart from some technical objections the question was raised why a relation that could be used for prediction purposes should exist at all. After all, a statistical relation between two sets of figures does not prove anything about the causal relation between the facts behind the sets of figures: one might cause the other, or reverse, or both may be caused by a third one. In the present case, the fate of a drug is determined by the bibliometric traits, or the bibliometric traits are determined by the clinical fate of a drug, or both are determined by some other factor. It is not possible to distinguish between the three possibilities on a statistical basis. In my opinion the bibliometric traits do not determine the fate of a drug, but the opposite is true. In that case the bibliometric traits will be apparent in the literature only after some time, and prediction on this basis is not possible. Thus, if Dr. Windsor wants to prove that such a prediction is possible, he has to prove that bibliometric traits cause the clinical success of a drug. I should be very interested to hear his arguments.

[1] D. A. Windsor, *J. Chem. Inf. Comput. Sci.*, **19**, 218–221 (1980).
[2] D. A. Windsor, "Could Bibliometric Data Be Used to Predict the Clinical Success of a Drug?", *J. Doc.*, **32** (3), 174–81 (1976).
[3] M. Osinga, *J. Doc.*, **33** (3), 239–40 (1977).

M. Osinga, **Editor**
Gist-Brocades nv.
Research & Development
P.O. Box 523, 2003 RM Haarlem, Holland

NEWS AND NOTES

NEWS ITEMS

TATE RECEIVES COLUMBUS AWARD

Dr. Fred A. Tate, associate director for planning and development at Chemical Abstracts Service, has received the 1980 Columbus Section Award of the American Chemical Society.

The award recognizes Dr. Tate's leadership in introducing computer-based information handling systems and procedures at Chemical Abstracts Service and his contributions in negotiating agreements under which organizations in the United Kingdom, West Germany, France, and Japan share in the financial support of CAS and the use of its data base. The award is sponsored by Ashland Chemical Co.

Dr. Tate received the Skolnik Award of the American Chemical Society's Division of Chemical Information in 1978.

CAS SEARCH ASSISTANCE DESK

Those who wish help or advice in searching Chemical Abstracts Service publications or computer files can now get it through the Search Assistance Desk at CAS.

The desk is staffed by chemists with a thorough knowledge of all CAS publications and files. They will answer questions about how CAS indexing terminology and practices affect searches for references on particular topics or substances in *Chemical Abstracts*, the *CA Search* computer-readable file (including the online chemical dictionary files), and other CAS publications and online computer files.

The Search Assistance Desk is in operation Monday through Friday from 8 a.m. to 5 p.m. Eastern time. It can be reached by calling (614) 421-6940, extension 3209. Written questions may be directed to Search Assistance Desk, Chemical Abstracts Service, P.O. Box 3012, Columbus, OH 43210.

ASIDIC MEETING

The Association of Information and Disseminination Centers (ASIDIC) will be sponsoring its 1980 fall meeting September 22 and 23 in Atlanta, Georgia. This year's conference will center on the theme "The Reuse of Information".

The conference, to be held at the Marriott Hotel, Courtland at International Blvd., Atlanta, will include talks by many recognized speakers on the pros and cons of this important topic. Working groups also will give attendees a chance to explore their own concerns.

For further details and registration information, contact: ASIDIC Secretariat, P.O. Box 8105, Athens, GA 30603, (404) 542-3106.