The Conversion of Wiswesser Line Notations to Ring Codes. I. The Conversion of Ring Systems*

CHARLES E. GRANITO** and SCOTT ROBERTS
Institute for Scientific Information
325 Chestnut St., Philadelphia, Pa. 19106

GERALD W. GIBSON
College of Charleston, Charleston, S.C. 29401

Received March 28, 1972

The computerized conversion of Wiswesser Line Notations to Ring Codes, using a two-part approach, and the set of computer programs generated for the conversion of ring systems are described.

In 1968, the Institute for Scientific Information (ISI) introduced its *Index Chemicus Registry System (ICRS),*[1] which utilizes Wiswesser Line Notations (WLN's) for processing the new chemical compounds being reported in *Current Abstracts of Chemistry and Index Chemicus (CAC&IC).* Over 150,000 new compounds are reported each year. *ICRS* is designed to provide chemists with current and retrospective chemical information via computer searches. To make *ICRS* useful to a wider audience, ISI has been investigating the conversion of WLN's to Ring Codes. This paper discusses just one part of the algorithm created to effect this conversion—i.e., the handling of ring structures (approximately 25% of the codes). A subsequent paper will cover the remaining Ring Codes.

BACKGROUND OF STRUCTURE HANDLING SYSTEMS

There are three basic chemical substructure handling systems: fragment codes, line notations, and connectivity tables. Up until about 1960, fragment codes were the only widely used system. After 1960, line notations, and somewhat later connectivity tables began to replace fragment codes.

The Wiswesser Line Notation (WLN),[2] being the best known chemical notation, was used by ISI in introducing its *Chemical Substructure Index* (CSI)[4] in 1971 (more than 1,000,000 permuted WLN entries per year). It is now used by about 100 major organizations around the world, and because of its many advantages,[3] it is being adopted by more organizations every year.

INTERCONVERSION OF SYSTEMS

Several WLN/CT interconversion studies have been made,[5] mainly because *Chemical Abstracts* uses connectivity tables (CT's)[6] in its registry system. The conversion of fragment codes to WLN's or CT's is impossible for most systems because the interrelationship of the fragments are usually not expressed. And although the converse is possible—i.e., conversion of line notations or CT's to the well established fragment codes—it, unfortunately, has received very little attention.

Because there are attractive alternatives (such as the WLN), most one-organization systems will probably be abandoned. However, a few fragment codes have been carefully tailored by groups of users over many years and will continue to survive. One of these is the Ring Code System.

RING CODES

Although used by about 100 organizations around the world, very little has been published in the open literature on the Ring Codes.[7] Ring Codes are fragment codes (Nubling and Steidle[8]). Some 300 codes have been adopted for use; these codes correspond to specific fragments (e.g., an $-NO_2$ group) as well as quantities (e.g., four rings or two nitrogens in a heterocyclic system).

Figure 1 shows the Ring Codes used for general compounds. There are additional codes for steroids and peptides; these are also being investigated. The general codes cover about 94% of the new compounds being reported in *Current Abstracts of Chemistry.* In Figure 1, only the codes of columns 2 through 8 are covered in this paper. The numbers appearing at the top of each box in Figure 1 refer to punch card positions and are used here for referencing individual fragments.

WISWESSER LINE NOTATIONS (WLN's)

WLN's uniquely and unambiguously represent every detail of a chemical structure. And they contain all of the information defined by Ring Codes. Theoretically, all Ring Codes could be produced from WLN's. However, the conciseness of WLN and the complexity of some of the Ring Code definitions make a 100% conversion difficult.

*Presented in part before the Division of Chemical Literature, 163rd Meeting, ACS, September 1971, Washington, D.C.

**To whom inquiries should be addressed.

1/12 SPECIAL CODE	2/12 1	3/12 ISOLATED	4/12 ISOLATED	5/12 LINEAR	6/12 ISOLATED	7/12 POLY	8/12 POLY	9/12 POLY	10/12 POLY	11/12 POLY	12/12 POLY	13/12 POLY	14/12 POLY	15/12 POLY	16/12 LINEAR
1/11 STEROID	2/11 2	3/11 COND AROMAT.	4/11 COND ALICYCL.	5/11 ANGULAR	6/11 COND. HET. CYCL.	7/11 SEVERAL LIKE X IN ONE RING	8/11 X COMMON TO SEVERAL RINGS	9/11 ANGULAR SUBSTITUTED	10/11 HETEROCYCLIC	11/11 2	12/11 >CH3	13/11 -C-, -C-,CX,C=Y	14/11 1	15/11 CIS	16/11 CROSSED
1/0 CARBOHYDRATE	2/0 3	3/0 COND. ALICYCL.	4/0 COND. HET. CYCL.	5/0 ANG-ANG.	6/0	7/0 SEVERAL UNLIKE X IN ONE RING	8/0 α	9/0 GEMINAL SUBSTITUTED	10/0 HETEROSUBSTIT.	11/0 3	12/0 -CH	13/0 C-X, C-X	14/0 2	15/0 TRANS.	16/0 CHAIN CHAIN
1/1 PEPTIDE	2/1 4	3/1 COND. HET. CYCL.	4/1 1	5/1 PERI	6/1	7/1 1N	8/1 β	9/1 ALICYCLIC	10/1 ALICYCLIC	11/1 4	12/1 -C-	13/1 C-X, C-X'	14/1 3	15/1 $A_2C=CA_2$, $A>C=CA_2$	16/1 CHAIN RING
1/2 NUCLEIC ACIDS	2/2 5	3/2 1	4/2 2	5/2 BRIDGE	6/2	7/2 2N	8/2 1,2	9/2 α	10/2 α1	11/2 5/10	12/2 CH_3X	13/2 C-Xm C-Xm, C-Xn C-X'n	14/2 4/6	15/2 >C=CA2	16/2 CHAIN AROMATIC
1/3 OTHER NAT OCCURING SUBSTS	2/3 ≥6	3/3 2	4/3 ≥3	5/3 SPIRO	6/3	7/3 3N	8/3 1,3	9/3 β	10/3 β2	11/3 11/16	12/3 -CH2N (CHAIN)	13/3 C-X, C=Y	14/3 ≥7	15/3 A, A >C=C	16/3 RING RING
1/4 SYNTHETIC PLASTICS AND OTHER POLYMERS	2/4 ISOLATED FROM	3/4 3	4/4 1	5/4 7 WITH 3=	6/4 ≥2	7/4 ≥4N	8/4 1, ≥4	9/4 1,2	10/4 1,2 3	11/4 ≥17	12/4 -CH2X (CHAIN)	13/4 C-X, C-C	14/4	15/4 A >C=C	16/4 WITH ≥6 RING MEMBER
1/5	2/5 IDENTIFICATION	3/5 4	4/5 ≥2	5/5 ≥7 UNSAT.	6/5	7/5 10	8/5 ≥3 SYM	9/5 1,3	10/5 1,3 4	11/5 1	12/5 (>CH2 N, (>CH2 X	13/5	14/5 -C≡C-	15/5 >C=C<	16/5 C≡C
1/6 INORGANIC SUBSTANCE	2/6 PROPERTIES	3/6 ≥5	4/6	5/6 ≥7 SAT.	6/6	7/6 ≥20	8/6 ≥3 ASYM.	9/6 ≥1,4	10/6 ≥ 1,4 5	11/6 2	12/6 >CHX	13/6	14/6 -C≡CH	15/6 C=C EXO	16/6
1/7 CLEAR TEXT	2/7 SYNTHETIC PREPN.	3/7 ARYL	4/7	5/7	6/7	7/7 S	8/7 ≥3 VIC	9/7 1,2,3 1,3,5	10/7 1,2,3 1,3,5 6 VIC. SYM	11/7 3	12/7 >CX	13/7 C=Y C=Y, C=Y C-YH	14/7 -C≡CX	15/7 1 X	16/7 SATURATED ALKYL
1/8 PATENT	2/8 UNCERTAIN STRUCTURE	3/8 ARALKYL	4/8	5/8	6/8 ≥7	7/8	8/8 X=C	9/8 1,2,4	10/8 1,2,4 ASYM 7	11/8 4/6	12/8 CH_2X_2 -CHX_2 >CX_2	13/8 C=Y C=Y, C=Y C-YH	14/8 -C=C=C -C=C=Y	15/8 ≥2 X ($X_1 = X_2$)	16/8 UNSATURATED ALKYL
1/9 FROM THE LITERATURE	2/9 GENERAL FORMULA	3/9	4/9	5/9 ALICYCLIC	6/9	7/9 OTHER HETERO	8/9 HETEROCYCLIC	9/9 ≥4 SUBST	10/9 ≥4 ≥8 SUBST.	11/9 ≥7	12/9 CHX_3 -CX_3 CX_4	13/9	14/9	15/9 ≥2 X ($X_1 \neq X_2$)	16/9

17/12 POLY	18/12 POLY	19/12 POLY	20/12 POLY	21/12 POLY	22/12 POLY	23/12 POLY	24/12 POLY	25/12 POLY	26/12 H	27/12 D,T	Sp. 28	Sp. 29	30/12
17/11 SUBST AT =	18/11 SUBST AT =	19/11 SUBST AT =	20/11 SUBST AT =	21/11 H	22/11 SUBST AT =	23/11 SUBST AT =	24/11 SUBST AT =	25/11 SUBST AT =	26/11 R	27/11 C*			30/11
17/0 F	18/0 1 (4) OH	19/0 R-NH2	20/0 >S=Y	21/0 R	22/0 -N=N-	23/0 -C-C=Y, H	24/0 H	25/0 (>=Y	26/0 OH,OM,OR SH,SM,SR	27/0 OTHER RADIOACT			30/0
17/1 Cl	18/1 2 (5) OH	19/1 R-NHT	20/1 >S=Y, >S=Y	21/1 Sp 17 Sp. 20	22/1 =N=N	23/1 -C-C=Y, R	24/1 OH,OM,OT SH,SM,ST	25/1	26/1 NH_2	27/1 Na 1 Li			30/1
17/2 Br	18/2 3 (≥6) OH	19/2 R-NHR	20/2 >P-	21/2 OH	22/2 -N=N+	23/2 -C, OH(SH)	24/2 OR, SR	25/2	26/2 NHR	27/2 Mg Hg 2			30/2
17/3 I	18/3 -OCH_3	19/3 R-NT_2	20/3 >Si<, R	21/3 OR	22/3 CHAIN	23/3 -C, OM(SM OT, ST)	24/3 NT_2	25/3 Y=Y	26/3 NR_2	27/3 Al B 3			30/3
17/4 NO_2	18/4 -OR	19/4 R-NRT	20/4 >Si<	21/4 SH,SR	22/4 NT_2	23/4 -C, OR(SR)	24/4 =NT	25/4	26/4 OX,NX,SX	27/4 Si 4			30/4
17/5 NO	18/5 -ONT	19/5 R-NR_2	20/5 B	21/5 OM,SM OT,ST	22/5 OH,OM OR,OT	23/5 -C, NT_2	24/5 Sp 17	25/5	26/5 Sp. 17,20	27/5 P 5			30/5
17/6 C≡N	18/6 ONIUM	19/6 R-$N(CH_3)_2$	20/6 SUBST IN RING	21/6 NT_2	22/6 OTHER HETERO	23/6 -C-C-X=Sp, X 17,20 ANHYDRIDES PERACIDS	24/6 Sp 20	25/6	26/6 R^+X^-	27/6 S 6			30/6 PATTERN
17/7 N=C	18/7 S	19/7 +$-NR_3$	20/7 Y=O	21/7 -OP--SP, -OS--SS-	22/7 N→O	23/7 Y=O	24/7 Y=O	25/7 Y=O	26/7 R^+R^-	27/7 HAL 7			30/7 TRACE ELEMENTS
17/8 SUBST AT (X	18/8 O/S	19/8 N+ CYCLIC	20/8 Y=S	21/8 -O-Si	22/8 OTHER	23/8 Y=S	24/8 Y=S	25/8 Y=S	26/8 COMPLEX ADDITION COMPOUNDS	27/8 Fe,Ni 8 CO			30/8 ACTIVE SUBSTANCE
17/9 SUBST GENERAL	18/9 HOMOLOGS	19/9 OZONIDE AZIDE	20/9 Y=NT	21/9 HOMOLOGS	22/9 HOMOLOGS	23/9 Y=NT	24/9 Y=NT	25/9 Y=NT	26/9 ORGANOMETALLIC RADICAL	27/9 OTHER			30/9 SUBSTANCE ACTED UPON

Figure 1. Ring Codes for general compounds

CONNECTIVITY TABLES

Although some of the Ring Code fragments could be obtained from a simple character-by-character string search of WLN (e.g., the presence of functional groups, e.g., OH); or specific rings (e.g., benzene), others require a rather intensive examination of the WLN symbols (e.g., those fragments that could be part of a ring *and* part of a substituent). To simplify the task, all WLN's are expanded to a connectivity table. Rather than working with one of the existing CT's, a new one was created. This decision was based partly on programming languages and degree of completeness of existing systems, and our desire to utilize a computer limited to 64K (bytes).

This new CT program expands WLN's by using Wiswesser dot-plot symbols as suggested by Hyde, Matthews, *et al.*[9] However, to meet our special needs a slightly different approach was taken. Figure 2 is an example of the ring records stored.

RING SYSTEMS

As we have mentioned, this paper discusses only the conversion of ring systems. In WLN, all ring systems except benzene are coded within L .... J or T .... J "brackets." Carbocyclic ring descriptions start with an L; heterocyclics start with a T. When an L or T starts a notation, a ring handling subroutine is activated. The following is a brief summary of how the computer algorithm handles various ring systems.

GENERAL APPROACH

For converting WLN's to Ring Codes, a number of registers are set up in the computer. As a WLN is read and processed, information is extracted and stored in the appropriate register(s). Additional registers are built up by combining information from the WLN and existing registers.

Eventually, both the topology and make-up of the compound are contained in an assortment of registers and are used in the conversion to appropriate Ring Codes.

For converting ring systems, the approach is to build up four registers: an Alphabetic Sequence Register (AS), Ring Unit Sequence Register (RUS), Composite Ring Register (CR), and Composite Ring Unit Register (CRU). These are used to show (1) the locants' number (AS) and make-up (RUS) in *alphabetic sequence* and (2) the identity (CR) and make-up (CRU) of the locants in the individual rings of a larger system, as they are connected. These are then used in making decisions about the molecule that is not directly available from the WLN.

A number of other registers are needed to count things asked for in the Ring Codes, as well as for referring, on demand, to bridges, multicyclic points, pseudo bridges, etc. The main conversion registers are shown in Figure 3.

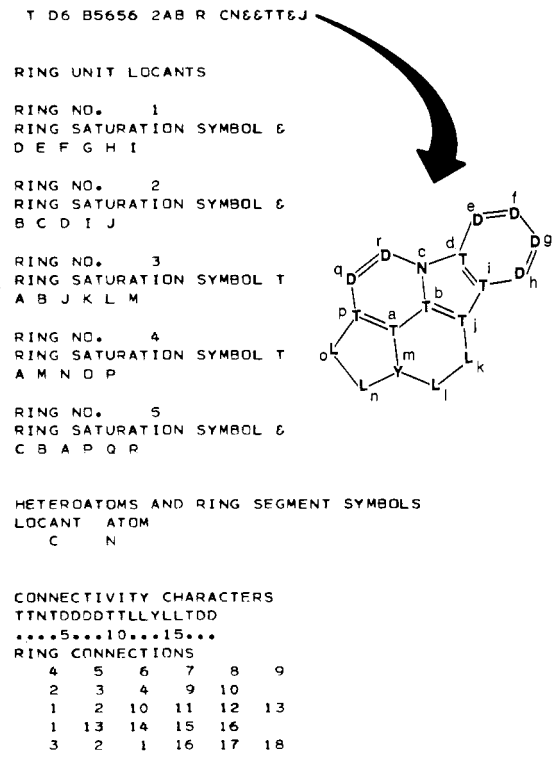


Figure 2. Stored ring records

RING SYSTEM APPROACH FOR CONVERSION OF WLN's TO CT's

To explain the over-all approach for converting WLN's to CT's the following examples are given. In each case, the order of connection for each ring is established as it is encountered in the WLN. Shored sides (fusion junctions) will lead to rings with common connection pairs.

Example No. 1: *Simple Polycyclic Fused Ring Systems*

Procedure: (1) Beginning with the first cited ring locant, list the following letters in the alphabet until Σ letters = R (R = ring numerals).

(2) For the next ring, begin with the cited (or understood for A) ring locant. List the following letters until the first letter of a *previous* ring is reached. Skip to final locant of previous ring and list the following letters until Σ letters = R_2.

(3) For subsequent rings repeat this process (Figure 4).

Example No. 2: *Perifused Ring Systems*

Procedure: (1) Proceed as with simple systems for the first ring.

(2) Proceed as with simple systems for subsequent rings, *unless* a cited ring locant is also a multicyclic point.

(3) If a cited ring locant is a multicyclic point, then, for the *third* ring in which it appears, precede the ring locant with the letter following it in the alphabet. If *that* letter is also a multicyclic point which has appeared in two previous rings, continue this "reverse alphabet" approach until a non-multicyclic point locant is reached. Next, move on to the first locant of the previous ring with the same earliest locant and continue as before until Σ letters = Rn (Figure 5).

Example No. 3: *Polycyclic Bridged Ring Systems*

Procedure: Same "reverse alphabet" approach, except that it is used the *second* time the bridge locant appears in a ring (Figures 6 and 7).

The procedures discussed above permit creation of CT's from WLN's. However, additional information is stored (e.g., individual ring descriptions) in anticipation of the Ring Code conversion.

(AS) ALPHABETIC SEQUENCE—locants of ring system in alphabetic order

(RUS) RING UNIT SEQUENCE—symbols for ring units in same order as (AS)

(CR) COMPOSITE RING—locants of rings as connected in an isolated or larger system

(CRU) COMPOSITE RING UNIT—symbols for ring units in same order as (CR)

(MA) MASTER ALPHABET—list of WLN alphabet in order

(RFS) RING FUSION & SIZE—list of earliest fusion locant and ring size of each ring in a larger system

(RN) RING NUMERAL—ring numeral of ring being considered, subtracted from as locants added to (CR)

(PB) PSEUDOBRIDGE—locants of atoms in pseudobridge

(MP) MULTICYCLIC POINT—locants of multicyclic points

(BA) BRIDGE ATOM—locants of bridge atoms

(BL) BRANCHED LOCANT—locants which are branched

(LI) LOCANT IDENTITY—holds locants until decision is made whether they are followed by & or -, etc.

(SR) SATURATED RINGS—lists T for saturated, and for unsaturated rings in order

(LL) LAST LOCANT—identifies last locant

(HL) HETEROLOCANT—locants and symbols for heteroatoms, X-points, etc. in a ring system

(RDE) RINGCODE ENTRY—"col./punch" corresponding to Ringcode code

(UH) U-HOLD—holds locants for doubly-bonded atoms cited in WLN or implied

(UUH) UU-HOLD—holds locants for triply-bonded atoms cited or implied

(RUSH) (RUS)-HOLD—holds (RUS) symbols for "insertion" of -MT-, H, W.

(HH) H-HOLD—holds locants cited with a saturation "H"

(GH) GENERAL HOLD—holding register for general use

(UC) UNSATURATED CARBON—holds number of unsaturated carbons encountered in check for double bonds

(UCH) (UC)-HOLD—holds (UC) total temporarily

(LTR) L OR T RING—listing in order of carbocyclic or heterocyclic nature of rings

(BAH) BENZENE, ALICYCLE, HETEROCYCLE—list in order of whether ring is B, A, H

(SRA) SHARED RING ATOM—locants of hetero atoms shared by 2 or more rings

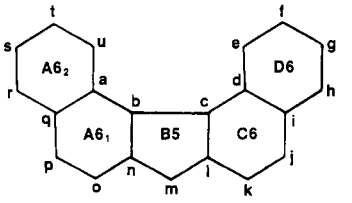
(SL) SUBSTITUENT LOCANT—locants of ring substituents

(OH) OTHER HETEROATOMS—list of heteroatoms other than N, O, S in a ring

Counting Registers:

(TRC) TOTAL RING COUNT
(CRC) CONDENSED RING COUNT
(BRC) BENZENE RING COUNT
(A6C) ALICYCLE-6 COUNT
(A5C) ALICYCLE-5 COUNT
(A7C) ALICYCLE-7 COUNT
(H6C) HETEROCYCLE-6 COUNT
(H5C) HETEROCYCLE-5 COUNT
(DBC) DOUBLE BOND COUNT
(NC) NITROGEN COUNT
(OC) OXYGEN COUNT
(SC) SULFUR COUNT
(A6SC) ALICYCLE-6 (Sat.) COUNT
(A61U) ALICYCLE-6, 1 UNSATURATION
(A62U) ALICYCLE-6, 2 UNSATURATIONS
(H61U) HETEROCYCLE-6, 1 UNSATURATION

Figure 3. Main conversion registers

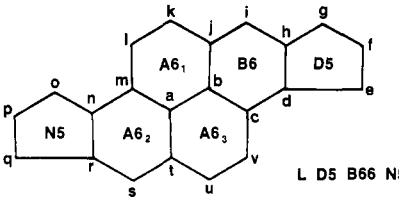


L D6 C6 B566TJ

Ring	Locants	Nonconsecutive Connections
D6	...defghi.....	d-i
C6	..cd....ijkL.....	d-i c-L
B5	.bc.....Lmn.....	c-L b-n
$A6_1$	ab.....nopq....	b-n a-q
$A6_2$	a.....qrstu	a-q a-u

Composite: defghi/cdijkL/bcLmn/abnopq/aqrstu

Figure 4. Simple polycyclic fused ring systems



L D5 B66 N566 2AB VTJ

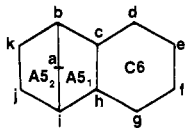
Ring	Locants	Nonconsecutive Connections
D5	...defgh.....	d-h
B6	.bcd..hij.....	d-h b-j
$A6_1$	ab.....jkLm.....	b-j a-m
N5	.....nopqr....	n-r
$A6_2$	a.....mn...rst..	a-m n-r a-t
$A6_3$	abc.....tuv cba	a-t c-v

Composite: defgh/bcdhij/abjkLm/nopqr/amnrst/cbatuv

Figure 5. Perifused ring systems

NUMBER OF RINGS IN TOTAL MOLECULE AND RING SYSTEM(s)

In Figure 1, the first six codes, 2/12 through 2/3, deal with ring counts. This information is directly available from the counting registers (Figure 3) and presents no problems. Codes 2/4 through 2/9 are not applicable because they deal with non-structural information or incompletely defined compounds (only defined compounds are registered in *ICRS*).



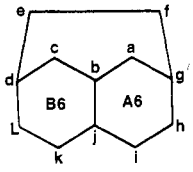
L C655 AJ

Ring	Locants	Nonconsecutive Connections
C6	. . cdefgh . . .	c-h
$A5_1$	abc . . . . hi . .	c-h a-i
$A5_2$	ab . . . . . ijk	a-i b-k

bridge

Composite: cdefgh/abchi/baijk

Figure 6. Polycyclic bridged ring systems



L76 B6 A C 1B LJ

Ring	Locants	Nonconsecutive Connections
A7	abcdefg . . .	a-g
A6	ab . . . ghij . .	a-g b-j
B6	bcd . . . jkL	b-j d-L

bridge

multicyclic point

Composite: abcdefg/baghij/dcbjkL

Figure 7. Polycyclic with both bridge and multicyclic point

BENZENE RINGS

Ring Codes deal with both isolated and fused benzene rings (Codes 3/12 through 3/6 in Figure 1). Isolated benzene rings are easily identified in WLN because the letter R is used for this purpose. However, for fused benzene rings the CT must be examined. Because each ring cited in WLN has been treated as a separate unit, the required information is available in the individual ring registers. Each "&" ring made up solely of D and/or T carbon atoms (D = —CH=, T = >C=) is recorded as a benzene ring. Because they refer to partly defined compounds, Codes 3/7 and 3/8 are not applicable. Code 3/9 has not been assigned.

ALICYCLIC RINGS

In WLN, alicyclics are identified as L ring systems. Individual ring sizes and counts are obtained from the registers shown in Figure 3. However, heterocyclic rings must also be examined to see if they contain any all-carbon rings fused to heterocyclic rings. Unsaturated carbons are noted in the UC register. This information is obtained from cited U's in WLN and by D(T) combinations in the CT registers when two or more alicyclic rings are present. Codes 4/12 through 4/9 and 5/4 through 5/8 are all readily handled. Code 2/11 uses heterocyclic information as discussed below.

RING LINKAGES (CODES 5/12 THROUGH 5/2)

Perifused rings (5/1) and bridged compounds (5/2) are identified by information directly available in WLN. The occurrence of a space-numeral-letter combination within the L . . . J or T . . . J brackets identifies perifused rings. The occurrence of the space-letter-space combination within L . . . J or T . . . J brackets identifies bridges.

Codes 5/12 through 5/$\emptyset$ are somewhat more troublesome. Here, the program must examine individual rings as recorded in CT with dot-plot symbols. These three codes are only possible for ring systems containing three or more fused and/or spiro linked rings. If the sequence of letters for locants in a given ring is not continuous and not evenly divided into two sets, then angular fusion occurs (5/11). If this occurs more than once, then 5/$\emptyset$ replaces 5/11 in the Ring Codes. The letter "a" is considered to follow the last letter used for this examination. The absence of angular fusion for three or more rings in one system produces the linear fusion (Code 5/12).

Spiro rings (X in WLN within T . . . J or L . . . J) require consideration of the X (spiro) points as well as the letter sequences. Here the consideration is also whether the sequence before an X point equals (in terms of number of symbols) the sequence after.

Multiple angular fusion can only occur in four or more rings. Examples of this approach are shown in Figure 8.

In cases like Example VI in Figure 8, the approach taken will not produce both 5/11 and 5/$\emptyset$ as called for by the present Ring Code definitions. However, this is not considered to be a serious deficiency of the system.

HETEROCYCLIC RINGS

Codes 6/12 through 6/9 and 7/12 through 7/9 do not present serious problems. However, Keto-enol isomerism is a problem: the Ring Code rules call for "enolizing" heterocyclics when possible. This requires converting a V (carbonyl) ring segment to a T-Q and changing an adjacent symbol to one with one less H implied—e.g., an NH to N or L to D. This conversion is required prior to determining whether codes such as 6/1 can be applied.

In the case of heteroatom positions (column 8), 8/12 and 8/11 are straightforward. Codes 8/$\emptyset$ and 8/1 require comparison of the locant of the heteroatom with the two end locants of the ring (e.g., a and f for an abcdef ring). If the heteroatom locant is *one* character later (or earlier) than the ring junction locant, it receives an 8/$\emptyset$; if not, an 8/1 is applied.

Codes 8/2 through 8/4 apply to two heteroatoms in a ring. If they are adjacent, an 8/2 applies. If they have one position separating them, an 8/3 applies. If they have two positions separating them, an 8/4 applies. These codes are obtained by examining the CT. If more than two heteroatoms are present, 8/5 through 8/7 must be considered. If all the heteroatoms are in sequence, 8/7

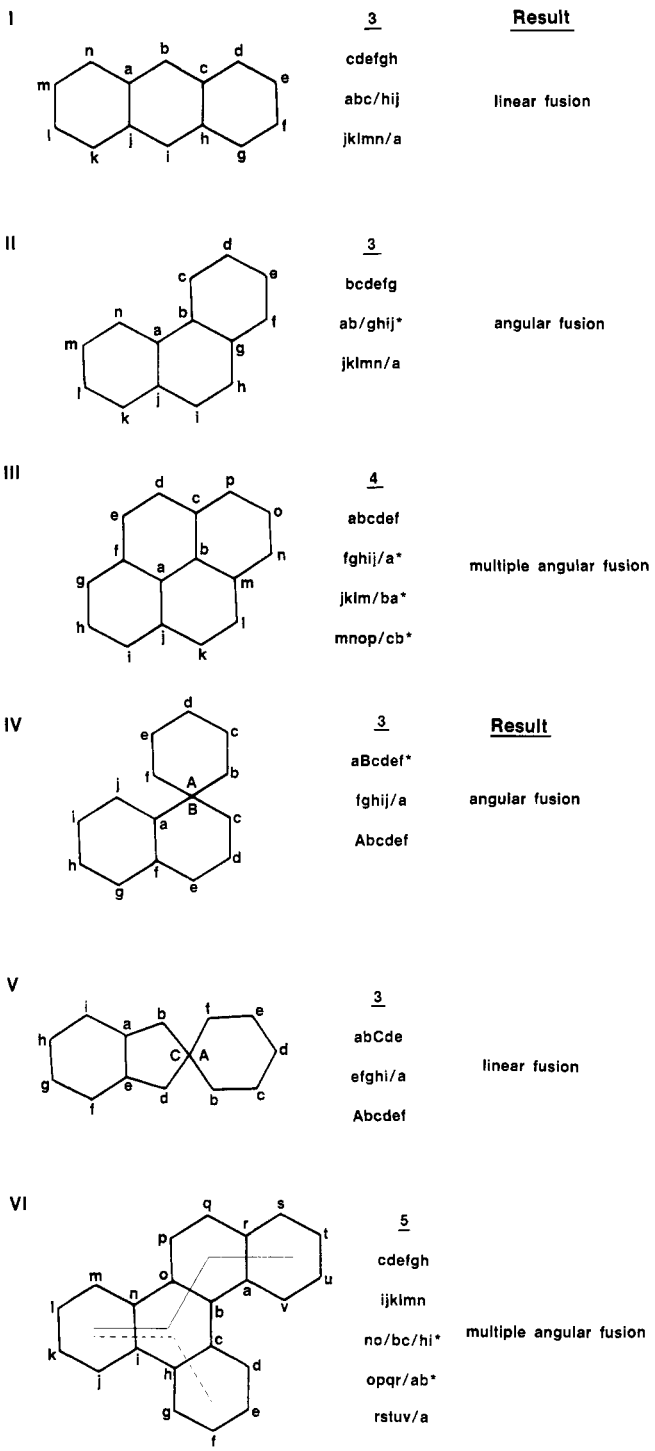


Figure 8. Linear angular fusions

applies. If the heteroatoms are not in sequence either 8/5 or 8/6 must apply.

The algorithm being discussed does not consider rings larger than 8 members for 8/5 or 8/6 punches.

For 6-, 7-, and 8-membered rings, the heteroatom count is compared to the ring size. If the difference is two, then 8/5 applies. If the difference is three or more, then specific asymmetrical arrangements are tested for (there are fewer asymmetrical than symmetrical cases). Figure 9 shows the arrangements by ring size and heteroatom. If the patterns tested fail, an 8/5 is applied.

	Hetero Atoms		
Ring Size	3	4	5
6	1,2,4	—	—
7	1,2,4	1,2,3,5	—
8	1,2,4	1,2,3,5	1,2,3,4,6
	1,2,5	1,2,4,6	1,2,3,5,6

Figure 9. Asymmetrical Patterns for Code 8/6

Rings having more than eight members (rarely encountered) are not covered in the algorithm being discussed. However, if considered necessary, the patterns shown in Figure 9 could be extended. In treating the CT record, V's, X's, and Y's are not considered heteroatoms.

Code 8/8 applies when a heteroatom is directly attached to a D or T in the CT, and the ring being considered is not fully unsaturated ("&" ring).

COMPUTER PROGRAM

The algorithm discussed has been incorporated into a computer program that is presently being tested at ISI.[10] The programs are written in Assembly language for an IBM 360. Running times will be discussed in the next paper (after large numbers of compounds have been processed).

CONCLUSIONS

A study concerning generation of all Ring Codes has been completed at ISI. The CT computer program now covers substituents and noncyclic compounds. Contraction and multiplier expansions have been worked out on paper and are being considered by a programmer. The remaining computer programs are scheduled for completion in 1972. A subsequent paper will deal with WLN/Ring Code conversion for the remaining Ring Codes.

The programs discussed in this paper should lead to valuable by-products, such as improved WLN checker programs and atom-by-atom searches for the small number of substructure questions (1 to 5%) where they would be useful.

Greater flexibility in substructure searching is the main advantage of having both WLN's and Ring Codes available. Also, the Ring Codes can be used as additional screens prior to searching directly on the WLN's. This should make the RADIICAL programs even more useful.

LITERATURE CITED

(1) Garfield, E., Revesz, G. S., Granito, C. E., Dorr, H. A., Calderon, M. M., and Warner, A., "Index Chemicus Registry System: Pragmatic Approach to Substructure Chemical Retrieval," *J. Chem. Doc.* **10** (1), 54-8 (1970).

(2) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968.

(3) Gelberg, A., "Chemical Notations," in "Encyclopedia of Library and Information Science," Vol. 4, pp. 510-28 Marcel Decker, New York, N.Y., 1970.

(4) Granito, C. E., and Rosenberg, M. D., "The Chemical Substructure Index (CSI) A New Research Tool," *J. Chem. Doc.* **11**, 251-6 (1971).

(5) Heller, S. R., and Koniver, D. A., "Computer Generation of Wiswesser Line Notation. II. Polyfused, Perifused, and Chained Ring Systems," *Ibid.*, 12 (1), 55-9 (1972).
(6) Leiter, D. P., Morgan, H. L., and Stobaugh, R. E., "Installation and Operation of a Registry for Chemical Compounds," *Ibid.*, 5, 238-42 (1965).
(7) Oatfield, H., "The ARCS System: Ringdoc As Used with a Computer," *Ibid.*, 7, 37-43 (1967).
(8) Nubling W., and Steidle, W., "The 'Documentationsring der Chemischpharmazeutischen Industrie'; Aims and Methods," *Angew. Chem. Internat. Edit.* 9 596-8 (1970).
(9) Hyde, E., Matthews, F. W., Thomson, L. H., and Wiswesser, W. J., "Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds," *J. Chem. Doc.* 7, 200-4 (1967).

Conceptual Design of an Information System for Toxicological Research and a Partial Implementation*

STEPHEN J. TAUBER,** HERBERT H. DER GIRAGOSSIAN, HAROLD S. CORBIN, JASPER P. HENDREN, and WILLIAM J. LaMARSH†
Informatics Inc., 6000 Executive Blvd., Rockville, Md. 20852

Received April 10, 1972

A design is described for a system to manage the information directly related to the experimental work in a toxicological laboratory. A number of distinct input procedures provide data to a set of co-existing data bases. Periodic, event-triggered, predefined demand, and ad hoc reports are posited. Data must be collected under stringent quarantine conditions. An initial implementation follows animals during experiments for weight, feed consumption, symptoms, and mode of death. A minicomputer, programmed in assembly language, controls automated source data collection. A larger computer, programmed in MARK IV, maintains the master file and produces the reports. The reports cover current experimental events, experimental status, time trends, and detailed data.

This work results from addressing the requirements of a toxicological research laboratory intended primarily for large-scale chronic studies of the effects of very small doses of potential toxicants on experimental animals.

In the first part of this paper we address the conceptual design of an information system to support animal experiments. The implementation of a subsystem is described in the second part of the paper. The information needs which the design seeks to satisfy were presented earlier.[1]

CONCEPTUAL DESIGN

The spectrum of information requirements of the toxicological laboratory, we allocate to three groups: access to existing external technical information, management of information directly pertaining to the experimental work, and management of ancillary information.

Access to external information involves considerations such as maintenance of a technical library and use of public and proprietary search systems, extensively discussed by many authors, including very recently such as Smith *et al.*,[2] Maizell *et al.*,[3] and Reinke.[4] Also included are who's-doing-what files such as are maintained by the Smithsonian Science Information Exchange on a continuing basis[5] and were compiled ad hoc by Vasta.[6]

Ancillary information is that which is only marginally affected by the specific nature of the work—e.g., inventory control and purchasing, building and equipment maintenance records, financial records, and personnel records. Some of these areas scarcely differ from those of other enterprises comparable in size. More peculiar are the special inventory control needed over sterilized supplies and the distinctive shipping procedures for purchased animals.

We restrict our consideration here to management of the information directly related to the experimental work. This encompasses direct experimental data, animal production, colony management, sample control, and scheduling. The design philosophy is to maintain distinct data bases for these several types of information, to use concerted access to them for output, and to distribute information to them from single input procedures in the laboratory.

Input. A protocol, which describes an experiment in detail, governs the laboratory activities. Each activity consists of a specific laboratory procedure with a pre-defined schedule. Changes may be made in the experiment based on preliminary data, reflected by changes in the procedures to be performed.

Data derive from the procedures and give rise to input transactions to the files. In some instances validation of data can be in real time even with batch updating. A skeletal view of the data flow from the major procedures into the data bases is given in the upper portion of Figure 1. The "Edit and Store" function must interpret the input record to determine the files affected, must provide appropriately formatted transactions for each, and must compute values derived partly from prior data.

*Presented at Division of Chemical Literature, 163rd Meeting, ACS, Boston, Mass., April 12, 1972.
**To whom correspondence should be addressed.
†Present address: Defense Communications Agency, Reston, Va. 22070.