Plausibility Check of Chemical Coding by Computer Using a Modified Derwent Robins Program

J. FITTING, H. LEHNA, and K. SPECHT*

Research Laboratories, Schering AG, Berlin/Bergkamen, Germany

Received December 27, 1973

The checking of chemical coding (Ringcode) by computer is described. The check is based on the normal search procedure for Ringcode coded files. Therefore the different logical conditions can be changed easily. Adoption to every chemical Ringcode is possible.

The plausibility check gives us the automatic possibility of finding errors in chemical coding. The coders get alert information about type and frequency of their own mistakes. This leads to optimal coding work. Time and costs consumed in manual checks of the coding quality are reduced to spot checks.

The main reasons for incorrect files are

1. Wrong interpretation of publications to be coded
2. Punching errors
3. Coding errors

The wrong interpretation of publications cannot be checked by computer, but, in this area, spot checks are sufficient to prove the standard of a coder.

Punching errors which occur nonsystematically can be found to a certain extent, but with an obligatory double check of the keypunching this type of error will be very rare.

Pure coding errors can be discovered by computer because of the logical connection of the different fragments. This detection will of course not be total, but a considerable number of mistakes will be found.

At first, the different logical correlations between the fragments of the chemical codes are determined. There are two possibilities to proceed:

1. to write a program which considers these correlations to a certain extent
2. to use a retrieval program and to define the correlations similar to the procedure for computer search runs

We decided on the second possibility because to us this solution seemed to be more flexible and has the option to be transferable to every similar chemical fragmentation code used now and in the future.

We combined a certain sum of check queries. After having them tested and evaluated we combined

31 Check Queries for the General Chemical Code
15 Check Queries for the Steroid Code

for the plausibility checks.

To illustrate the plausibility check method, an example is given: Figure 1 shows the encoding of a chemical compound with the General Chemical Code. The coding is wrong; the position 7/11 is missing. Using the plausibility check for the General Chemical Code, an error message will be given by the computer:

SEVERAL X IN ONE RING

Simultaneously a card with the wrong coding will be punched; the corresponding file on the tape is obliterated.

The logical correlations for this mentioned error are as follows. The positions 8/2–8/7 define the positions of heteroatoms in rings with more than one heteroatom. Additionally the positions 7/11 or 7/ϕ must be punched. Therefore, we use the following check formula

$$\underbrace{(8/2 \vee 8/3 \vee 8/4 \vee 8/5 \vee 8/6 \vee 8/7)}_{A} \wedge \underbrace{(7/11 \vee 7/\phi)}_{B}$$

or

$$\text{if A, then B } (A \longrightarrow B)$$

On the other hand, if the fragments 7/11 and 7/ϕ are present, there must be statements in the positions 8/2–8/7, too. Therefore, the logical condition

$$\text{if B, then A } (B \longrightarrow A)$$

is correct, too. The condition is alternate.

If we use the following definition of a query

$$(A \wedge \bar{B}) \vee (\bar{A} \wedge B)$$

we shall get *no* error message if both, A and B, are present, and if both are absent. Considering the fact that very frequently coordinate coding terms are both missing, we found it necessary not to check only A and B alternately but to check additionally one or both by further check positions:

Condensed heterocycles, for example, have the punch positions

$$\underbrace{(3/1 \vee 4/\phi \vee 6/11)}_{C}$$

and always need statements in 8/ϕ–8/7 (A'). A' has two more possible positions than A.

The new logical condition is

$$\text{if C, then A' } (C \rightarrow A')$$

Because

$$A' = (A \vee 8/\phi \vee 8/1) = > A$$

it is not possible to detect quantitatively the combined missing of A and B. But the possibility of overlooking erroneous codings is remarkably limited. Generally we try to set up a very fine mesh when we define the check queries.

Figure 2 shows all the punch positions which may be checked using our question set with regard to the coding of the substance shown in Figure 1.

Some changes of the Derwent Robins Program had to be made. For example, it was necessary to check every card of a file. To identify the erroneous cards, the card's sequence numbers have been included in the address. Files which do not have numbered cards will be numbered by the pro-

* Author to whom correspondence should be addressed.

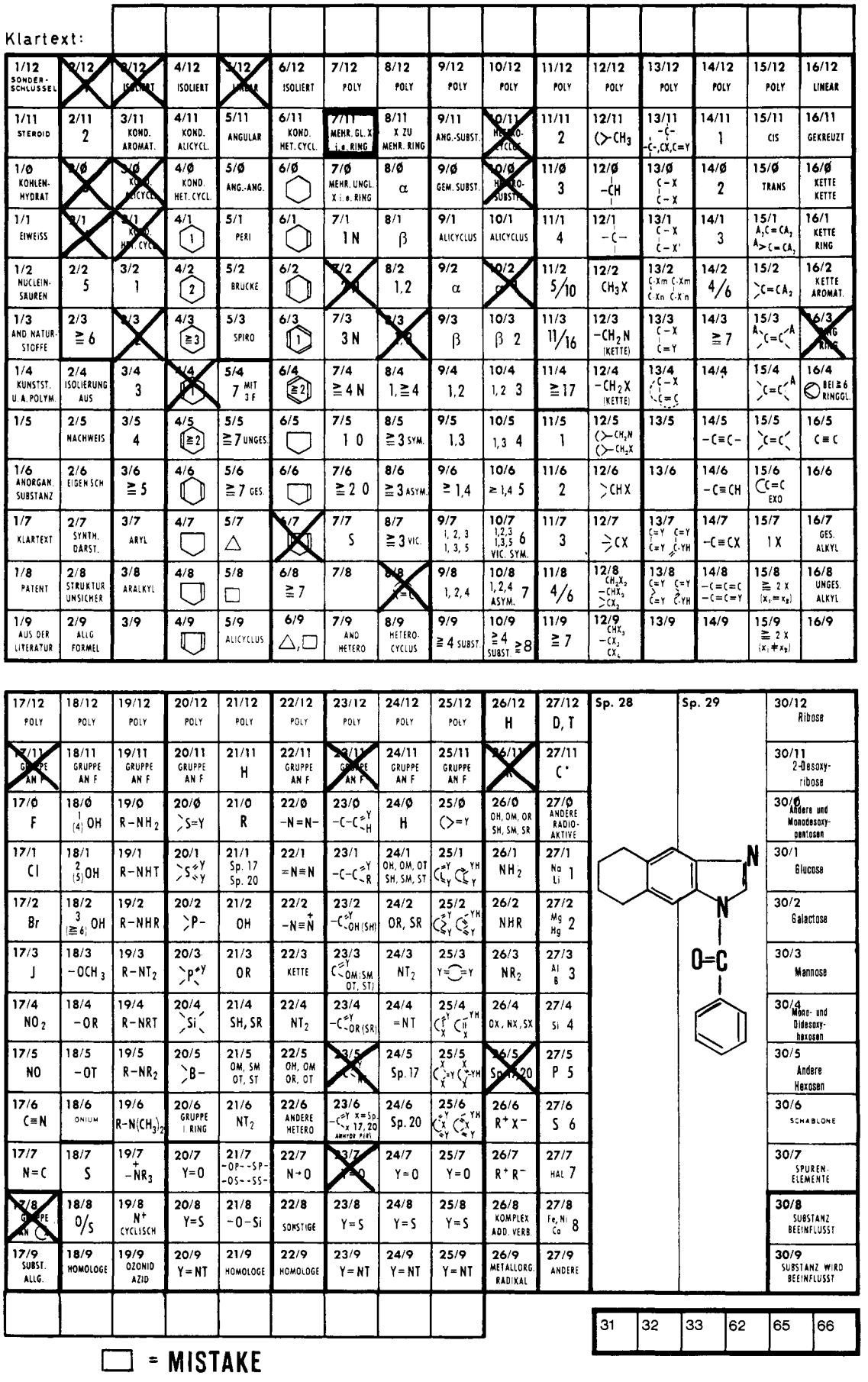

Klartext:
1/12 SONDER-SCHLÜSSEL
2/12
3/12 ISOLIERT
4/12 ISOLIERT
5/12 LINEAR
6/12 ISOLIERT
7/12 POLY
8/12 POLY
9/12 POLY
10/12 POLY
11/12 POLY
12/12 POLY
13/12 POLY
14/12 POLY
15/12 POLY
16/12 LINEAR
1/11 STEROID
2/11 2
3/11 KOND. AROMAT.
4/11 KOND. ALICYCL.
5/11 ANGULAR
6/11 KOND. HET. CYCL.
7/11 MEHR. GL. X i. e. RING
8/11 X ZU MEHR. RING
9/11 ANG.-SUBST.
10/11 HETERO-CYCLUS
11/11 2
12/11 CH3
13/11 C, CX, C=Y
14/11 1
15/11 CIS
16/11 GEKREUZT
1/Ø KOHLEN-HYDRAT
2/Ø
3/Ø KOND. ALICYCL.
4/Ø KOND. HET. CYCL.
5/Ø ANG.-ANG.
6/Ø
7/Ø MEHR. UNGL. X i. e. RING
8/Ø α
9/Ø GEM. SUBST.
10/Ø HETERO-SUBSTIT.
11/Ø 3
12/Ø CH
13/Ø C-X, C-X
14/Ø 2
15/Ø TRANS
16/Ø KETTE KETTE
1/1 EIWEISS
2/1
3/1 KOND. HET. CYCL.
4/1 1
5/1 PERI
6/1
7/1 1 N
8/1 β
9/1 ALICYCLUS
10/1 ALICYCLUS
11/1 4
12/1 C
13/1 C-X, C-X'
14/1 3
15/1 A2C=CA2, A>C=CA2
16/1 KETTE RING
1/2 NUCLEIN-SÄUREN
2/2 5
3/2 1
4/2 2
5/2 BRÜCKE
6/2
7/2
8/2 1,2
9/2 α
10/2
11/2 5/10
12/2 CH3X
13/2 C-Xm C-Xm C-Xn C-X n
14/2 4/6
15/2 C=CA2
16/2 KETTE AROMAT.
1/3 AND. NATUR-STOFFE
2/3 ≧6
3/3
4/3 ≧3
5/3 SPIRO
6/3 1
7/3 3 N
8/3
9/3 β
10/3 β 2
11/3 11/16
12/3 CH2N (KETTE)
13/3 C-X, C=Y
14/3 ≧7
15/3 C=C
16/3 RING
1/4 KUNSTST. U. A. POLYM.
2/4 ISOLIERUNG AUS
3/4 3
4/4
5/4 7 MIT 3 F
6/4 ≧2
7/4 ≧4 N
8/4 1, ≧4
9/4 1,2
10/4 1,2 3
11/4 ≧17
12/4 CH2X (KETTE)
13/4 C-X, C=C
14/4
15/4 C=C
16/4 BEI ≧6 RINGGL.
1/5
2/5 NACHWEIS
3/5 4
4/5 ≧2
5/5 ≧7 UNGES.
6/5
7/5 1 0
8/5 ≧3 SYM.
9/5 1.3
10/5 1,3 4
11/5 1
12/5 CH2N, CH2X
13/5
14/5 -C≡C-
15/5 C=C
16/5 C≡C
1/6 ANORGAN. SUBSTANZ
2/6 EIGEN SCH
3/6 ≧5
4/6
5/6 ≧7 GES.
6/6
7/6 ≧2 0
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 KLARTEXT
2/7 SYNTH. DARST.
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 GES. ALKYL
1/8 PATENT
2/8 STRUKTUR UNSICHER
3/8 ARALKYL
4/8
5/8
6/8 ≧7
7/8
8/8
9/8 1, 2, 4
10/8 1,2,4 ASYM. 7
11/8 4/6
12/8 CH2X2, CHX3, CX4
13/8 C=Y, C=Y, C=Y, C-YH
14/8 -C=C=C, -C=C=Y
15/8 ≧ 2 X (X1=X2)
16/8 UNGES. ALKYL
1/9 AUS DER LITERATUR
2/9 ALLG FORMEL
3/9
4/9
5/9 ALICYCLUS
6/9 △, □
7/9 AND HETERO
8/9 HETERO-CYCLUS
9/9 ≧4 SUBST.
10/9 ≧4 SUBST. ≧8
11/9 ≧7
12/9 CHX3, CX3, CX4
13/9
14/9
15/9 ≧ 2 X (X1≠X2)
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 Ribose
17/11 GRUPPE AN F
18/11 GRUPPE AN F
19/11 GRUPPE AN F
20/11 GRUPPE AN F
21/11 H
22/11 GRUPPE AN F
23/11 GRUPPE AN F
24/11 GRUPPE AN F
25/11 GRUPPE AN F
26/11
27/11 C*
30/11 2-Desoxy-ribose
17/Ø F
18/Ø (4) OH
19/Ø R-NH2
20/Ø S=Y
21/Ø R
22/Ø -N=N-
23/Ø -C-C=Y, H
24/Ø H
25/Ø =Y
26/Ø OH, OM, OR SH, SM, SR
27/Ø ANDERE RADIO-AKTIVE
30/Ø Andere und Monodesoxy-pentosen
17/1 Cl
18/1 (5) OH
19/1 R-NHT
20/1 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 NH2
27/1 Na Li 1
30/1 Glucose
17/2 Br
18/2 (≧6) OH
19/2 R-NHR
20/2 P-
21/2 OH
22/2 -N≡N
23/2 -C=Y, OH(SH)
24/2 OR, SR
25/2
26/2 NHR
27/2 Mg Hg 2
30/2 Galactose
17/3 J
18/3 -OCH3
19/3 R-NT2
20/3 P*Y
21/3 OR
22/3 KETTE
23/3 C=Y, OM, SM, OT, ST
24/3 NT2
25/3 Y=, =Y
26/3 NR2
27/3 Al B 3
30/3 Mannose
17/4 NO2
18/4 -OR
19/4 R-NRT
20/4 Si
21/4 SH, SR
22/4 NT2
23/4 -C=Y, OR(SR)
24/4 =NT
25/4
26/4 OX, NX, SX
27/4 Si 4
30/4 Mono- und Didesoxy-hexosen
17/5 NO
18/5 -OT
19/5 R-NR2
20/5 B-
21/5 OM, SM OT, ST
22/5 OH, OM OR, OT
24/5 Sp. 17
25/5
26/5 Sp. 17, 20
27/5 P 5
30/5 Andere Hexosen
17/6 C≡N
18/6 ONIUM
19/6 R-N(CH3)2
20/6 GRUPPE I. RING
21/6 NT2
22/6 ANDERE HETERO
23/6 X=Sp. 17, 20 ARMYOR PERI
24/6 Sp. 20
25/6
26/6 R+X-
27/6 S 6
30/6 SCHABLONE
17/7 N=C
18/7 S
19/7 +NR3
20/7 Y=O
21/7 -OP-, -SP-, -OS-, -SS-
22/7 N→O
23/7
24/7 Y=O
25/7 Y=O
26/7 R+R-
27/7 HAL 7
30/7 SPUREN-ELEMENTE
17/8 GRUPPE AN
18/8 O/S
19/8 N+ CYCLISCH
20/8 Y=S
21/8 -O-Si
22/8 SONSTIGE
23/8 Y=S
24/8 Y=S
25/8 Y=S
26/8 KOMPLEX ADD. VERB.
27/8 Fe, Ni Co 8
30/8 SUBSTANZ BEEINFLUSST
17/9 SUBST. ALLG.
18/9 HOMOLOGE
19/9 OZONID AZID
20/9 Y=NT
21/9 HOMOLOGE
22/9 HOMOLOGE
23/9 Y=NT
24/9 Y=NT
25/9 Y=NT
26/9 METALLORG. RADIKAL
27/9 ANDERE
30/9 SUBSTANZ WIRD BEEINFLUSST
31
32
33
62
65
66
= MISTAKE
H3 BF XXIII 21253

Figure 1.

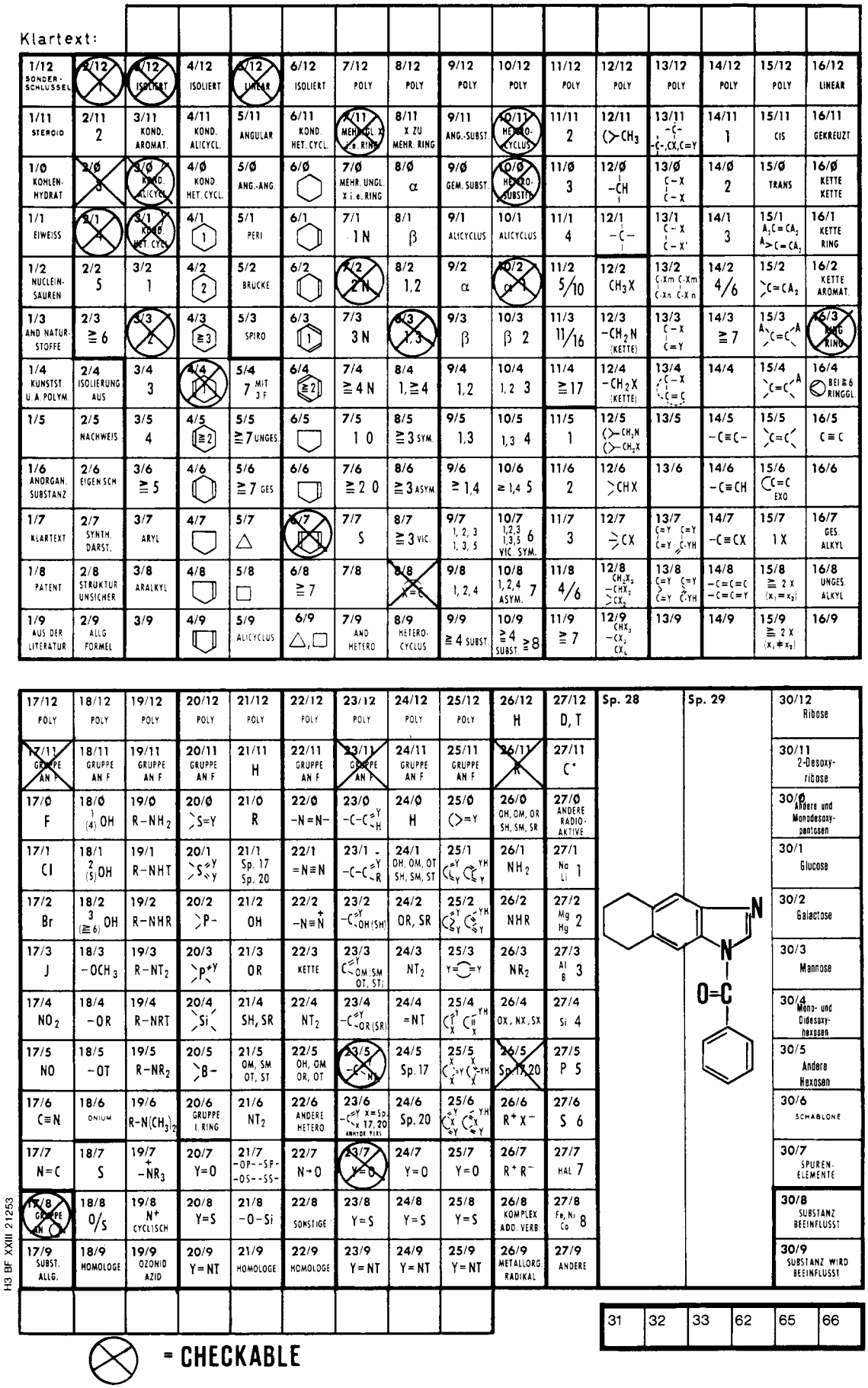

Klartext:
1/12 SONDER-SCHLUSSEL
2/12
3/12 ISOLIERT
4/12 ISOLIERT
5/12 LINEAR
6/12 ISOLIERT
7/12 POLY
8/12 POLY
9/12 POLY
10/12 POLY
11/12 POLY
12/12 POLY
13/12 POLY
14/12 POLY
15/12 POLY
16/12 LINEAR
1/11 STEROID
2/11 2
3/11 KOND. AROMAT.
4/11 KOND. ALICYCL.
5/11 ANGULAR
6/11 KOND. HET. CYCL.
7/11 MEHR. UNGL. X i. e. RING
8/11 X ZU MEHR. RING
9/11 ANG.-SUBST.
10/11 HETERO-CYCLUS
11/11 2
12/11 CH3
13/11
14/11 1
15/11 CIS
16/11 GEKREUZT
1/Ø KOHLEN-HYDRAT
2/Ø
3/Ø KOND. ALICYCL.
4/Ø KOND. HET. CYCL.
5/Ø ANG.-ANG.
6/Ø
7/Ø MEHR. UNGL. X i. e. RING
8/Ø α
9/Ø GEM. SUBST.
10/Ø HETERO-SUBSTIT.
11/Ø 3
12/Ø CH
13/Ø
14/Ø 2
15/Ø TRANS
16/Ø KETTE KETTE
1/1 EIWEISS
2/1
3/1 KOND. HET. CYCL.
4/1 1
5/1 PERI
6/1
7/1 1N
8/1 β
9/1 ALICYCLUS
10/1 ALICYCLUS
11/1 4
12/1
13/1
14/1 3
15/1 A2C=CA2 A2>C=CA2
16/1 KETTE RING
1/2 NUCLEIN-SAUREN
2/2 5
3/2 1
4/2 2
5/2 BRUCKE
6/2
7/2 2N
8/2 1,2
9/2 α
10/2 α 1
11/2 5/10
12/2 CH3X
13/2
14/2 4/6
15/2 >C=CA2
16/2 KETTE AROMAT.
1/3 AND NATUR-STOFFE
2/3 ≧6
3/3 2
4/3 ≧3
5/3 SPIRO
6/3 1
7/3 3N
8/3 1,3
9/3 β
10/3 β 2
11/3 11/16
12/3 -CH2N (KETTE)
13/3
14/3 ≧7
15/3
16/3 RING
1/4 KUNSTST. U. A. POLYM.
2/4 ISOLIERUNG AUS
3/4 3
4/4 1
5/4 7 MIT 3 F
6/4 ≧2
7/4 ≧4 N
8/4 1, ≧4
9/4 1,2
10/4 1,2 3
11/4 ≧17
12/4 -CH2X (KETTE)
13/4
14/4
15/4
16/4 BEI ≧6 RINGGL.
1/5
2/5 NACHWEIS
3/5 4
4/5 ≧2
5/5 ≧7 UNGES.
6/5
7/5 1 0
8/5 ≧3 SYM.
9/5 1,3
10/5 1,3 4
11/5 1
12/5
13/5
14/5 -C≡C-
15/5 >C=C<
16/5 C≡C
1/6 ANORGAN. SUBSTANZ
2/6 EIGEN SCH
3/6 ≧5
4/6
5/6 ≧7 GES
6/6
7/6 ≧2 0
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 KLARTEXT
2/7 SYNTH. DARST.
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
14/7 -C≡CX
15/7 1 X
16/7 GES. ALKYL
1/8 PATENT
2/8 STRUKTUR UNSICHER
3/8 ARALKYL
4/8
5/8 □
6/8 ≧7
7/8 S
8/8 X=C
9/8 1, 2, 4
10/8 1, 2, 4 ASYM. 7
11/8 4/6
12/8 CH2X2 -CHX2 >CX2
13/8
14/8 -C=C=C -C=C=Y
15/8 ≧ 2 X (X1=X2)
16/8 UNGES. ALKYL
1/9 AUS DER LITERATUR
2/9 ALLG FORMEL
3/9
4/9
5/9 ALICYCLUS
6/9 △, □
7/9 AND HETERO
8/9 HETERO-CYCLUS
9/9 ≧4 SUBST.
10/9 ≧4 SUBST. ≧8
11/9 ≧7
12/9 CHX3 -CX3 CX4
13/9
14/9
15/9 ≧ 2 X (X1≠X2)
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 Ribose
17/11 GRUPPE AN F
18/11 GRUPPE AN F
19/11 GRUPPE AN F
20/11 GRUPPE AN F
21/11 H
22/11 GRUPPE AN F
23/11 GRUPPE AN F
24/11 GRUPPE AN F
25/11 GRUPPE AN F
26/11 R
27/11 C*
30/11 2-Desoxy-ribose
17/Ø F
18/Ø (4) OH
19/Ø R-NH2
20/Ø >S=Y
21/Ø R
22/Ø -N=N-
23/Ø
24/Ø H
25/Ø
26/Ø OH, OM, OR SH, SM, SR
27/Ø ANDERE RADIO-AKTIVE
30/Ø Andere und Monodesoxy-pentosen
17/1 Cl
18/1 2 (5) OH
19/1 R-NHT
20/1
21/1 Sp. 17 Sp. 20
22/1 =N≡N
23/1
24/1 OH, OM, OT SH, SM, ST
25/1
26/1 NH2
27/1 Na Li 1
30/1 Glucose
17/2 Br
18/2 3 (≧6) OH
19/2 R-NHR
20/2 >P-
21/2 OH
22/2 -N≡N+
23/2
24/2 OR, SR
25/2
26/2 NHR
27/2 Mg Hg 2
30/2 Galactose
17/3 J
18/3 -OCH3
19/3 R-NT2
20/3
21/3 OR
22/3 KETTE
23/3
24/3 NT2
25/3
26/3 NR2
27/3 Al B 3
30/3 Mannose
17/4 NO2
18/4 -OR
19/4 R-NRT
20/4
21/4 SH, SR
22/4 NT2
23/4
24/4 =NT
25/4
26/4 OX, NX, SX
27/4 Si 4
30/4 Mono- und Didesoxy-hexosen
17/5 NO
18/5 -OT
19/5 R-NR2
20/5 >B-
21/5 OM, SM OT, ST
22/5 OH, OM OR, OT
23/5
24/5 Sp. 17
25/5
26/5 Sp. 17,20
27/5 P 5
30/5 Andere Hexosen
17/6 C≡N
18/6 ONIUM
19/6 R-N(CH3)2
20/6 GRUPPE I. RING
21/6 NT2
22/6 ANDERE HETERO
23/6 X=Sp. 17, 20 ANHYDR. PERS.
24/6 Sp. 20
25/6
26/6 R*X-
27/6 S 6
30/6 SCHABLONE
17/7 N=C
18/7 S
19/7 -NR3+
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 SPUREN-ELEMENTE
17/8 GRUPPE AN O
18/8 O/S
19/8 N+ CYCLISCH
20/8 Y=S
21/8 -O-Si
22/8 SONSTIGE
23/8 Y=S
24/8 Y=S
25/8 Y=S
26/8 KOMPLEX ADD. VERB
27/8 Fe, Ni Co 8
30/8 SUBSTANZ BEEINFLUSST
17/9 SUBST. ALLG.
18/9 HOMOLOGE
19/9 OZONID AZID
20/9 Y=NT
21/9 HOMOLOGE
22/9 HOMOLOGE
23/9 Y=NT
24/9 Y=NT
25/9 Y=NT
26/9 METALLORG. RADIKAL
27/9 ANDERE
30/9 SUBSTANZ WIRD BEEINFLUSST
31
32
33
62
65
66
H3 BF XXIII 21253
= CHECKABLE

Figure 2.

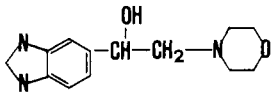
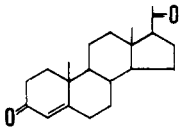
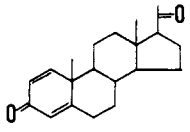
Coded compound	Punch positions erroneously absent	Output
	6/12 6/6	10082S 01 029 ISOLATED RING 014 UNSATURATED HETEROCYCLE
	2/1 25/11 25/Ø 25/7	10437S 01 035 SIDE CHAIN 044 COL. 25
	23/7	10437S 02 003 Y IN COL. 23

Figure 3.

gram. A new listprogram was written. This program allows the printing of accession number and card number of erroneous cards together with all mistakes (see Figure 3).

Contrary to a fixed check program, the procedure described here gives the possibility of changing the logical conditions very easily. Thus it is possible to consider the changing of coding rules a.s.o. without changing the program.

DRS—A User Oriented Information Retrieval System

GEZA SZONYI

Polaroid Corporation, Cambridge, Massachusetts 02139

Received December 6, 1973

DRS (Data Retrieval System), a user-oriented computerized information retrieval system, is described. The system allows data base generation and information retrieval by users without programming and systems knowledge. As an illustration, a chemical information system using the Wiswesser Line Notation (WLN) is generated, and retrieval examples are given using this data base. Because of its easily understandable English command language, DRS is suitable for a wide variety of information retrieval systems. Polaroid was instrumental in many of the advanced features of DRS developed during the past three years.

FAMULUS, a user-oriented computerized information retrieval system was presented in the fall of 1969.[1] This system, consisting of eight main and three peripheral computer programs, allows the user to set up data bases with relative ease. Having done-this, data manipulation, such as addition, correction, and deletion of information units or records, can be carried out relatively simply. Subsequently, searching of the data base by Boolean logic or by keywords is accomplished by English commands, and the retrieved subset can then be sorted alphabetically. In addition, data subunits or files can be merged, and printouts of the retrieved subset can be displayed in several output formats.[2] The system was mainly designed to handle personal reference collections but appeared to have applicability for all types of computerized information retrieval systems.[3] Unfortunately, the system was designed for IBM System/360 Control Data Corp. 6400 and 6600, UNIVAC 1108, or Xerox Sigma 7 type computers.[1,2] These are large machines and not particularly suited for operation by individual users without system and programming knowledge. Also, searches with this system are carried out in the batch mode, rather than interactively, making the information retrieval rather awkward. A more serious drawback of FAMULUS is its magnetic tape orientation; *i.e.*, searching is done sequentially rather than at random. In addition, typical data bases contain 3000–5000 records, each record having about 4000 characters.[2]

We were interested in FAMULUS-type retrieval systems, *i.e.*, having all the attributes of such a system, but one which could be implemented on minicomputers, since these could potentially be operated by single users without system or programming staff support. Further, data bases