## ACKNOWLEDGMENTS

The system described here developed over several years through the efforts of many people in Du Pont, especially the Central Report Index and Information Retrieval Systems groups. The author acknowledges the advice and help given by Dr. Melvin L. Huber of the Central Report Index in the preparation of this paper.

## LITERATURE CITED

(1) Gluck, D. J., "A Chemical Structure Storage and Search System Developed at Du Pont," *J. Chem. Doc.*, **5**, 43 (1965).
(2) Hoffman, W. S., "An Integrated Chemical Structure Storage and Search System Operating at Du Pont," *J. Chem. Doc.*, **8**, 3 (1968).
(3) Hoffman, W. S., "Du Pont Information Flow System," *J. Chem. Doc.*, **12**, 116 (1972).
(4) Leiter, D. P., Morgan, H. L., and Stobaugh, R. E., "Installation and Operation of a Registry for Chemical Compounds," *J. Chem. Doc.*, **5**, 238 (1965).
(5) Montague, B. A. and Schirmer, R. F., "Du Pont Central Report Index: System Design, Operation, and Performance," *J. Chem. Doc.*, **8**, 33 (1968).
(6) Morgan, H. L., "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service," *J. Chem. Doc.*, **5**, 107 (1965).

# Hoffmann-La Roche's On-Line/Batch Interactive Chemical Information System†

A. SHENG,* L. LUPI, M. RONAYNE, A. SPRULES, and S. ZORNETZER

Hoffmann-La Roche Inc., Nutley, New Jersey 07110

Received September 30, 1974

**This paper presents the current view of Roche's integrated Chemical Information System which is user-oriented and modularly designed for easy expansion. The essential elements of the data base for on-line interrogation and the various on-line search procedures—Ro number search and Wiswesser notation search—for different applications are described.**

## SYSTEM OVERVIEW

This paper provides an overview of Hoffmann-La Roche's Chemical Information System. It explains the main features of the system, the advantages and limitations of its various modules, and the means by which the subsystems are interrelated. The design characteristics of the system were modular, versatile, user-oriented, and open-ended to allow individual subsystems to be added to the overall integrated system for easy expansion and modification.

The system was jointly developed by the Research Systems Section of Management Information Services Department and the Research Records Office of Research Services Department. It is implemented on a Honeywell 600/6000 time-sharing system with CRTs and teletypewriters to form a network of terminal-to-computer communication on a data base of over 100,000 compounds. In addition, "Vistas," which are visual display devices monitored by a special software package, are placed in various locations to provide management and technical personnel the momentary status and utilization of the operating system. Each job is assigned an identifying name or number which is displayed on the Vista, thereby allowing the user to follow the step-by-step execution of his program (Figure 1).

The chemical information system provides internal information services for the scientists of the Research Division. Inquiry, search, and retrieval are performed on request through the Research Records Office (RRO) which serves as a focal point for storage and retrieval of technical information.

Figure 2 is a system flow chart showing the various technical components and the processing of the interrelated data elements within the chemical information system. The data bank consists of the following subsets: chemical name, chemical structure, chemist name, chemist number, molecular formula, and Wiswesser notation[1] of the registered compounds. Each compound is identified by a nine-digit number called the Ro number (Roche registry number). The three major files which form the essential components of the data base are the chemical name file, the chemical structure file, and the Wiswesser notation file. These are randomly structured files stored on magnetic disk and are processed by a series of programs written in Fortran and Assembly Language for man-machine interaction in entry, update, search, and retrieval.

## TERMINAL DEVICES AND INPUT PROCEDURE

**A. On-Line Terminals.** Two remote terminal devices, the teletypewriter and the CRT, have been used for communication with the central processor both to enter and retrieve information.

The teletype, Model 37, with paper tape punch and reader module, is commonly used for processing various subsystems. It is equipped with half forward and reverse line space, chemical bond symbols, and upper and lower case characters. A special feature of the character set is the inclusion of "Octobliques,"[2] an extension of the existing dot-bond notation, for depicting the spatial arrangement in complex structures. The set consists of eight oblique lines: two slants each of slope +2.0, −2.0, +0.5, and −0.5. Each pair of like slopes is situated in opposite halves of the character matrix.

Chemical data are handled by the following procedure. After a compound has been assigned an Ro number, the TTY 37 is used for recording the chemical data including the structure on a special form called a data sheet. The reverse line space feature allows the operator to type the structure along any convenient path, for example, around a

† Presented before the Division of Chemical Literature, 168th National Meeting of the American Chemical Society, Atlantic City, N.J., Sept 10, 1974.

* To whom correspondence should be addressed.

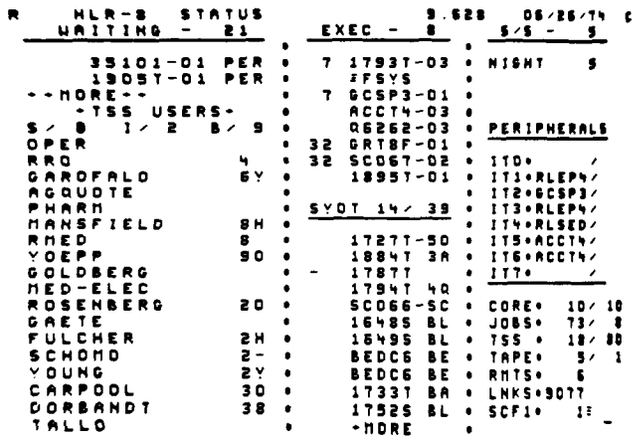
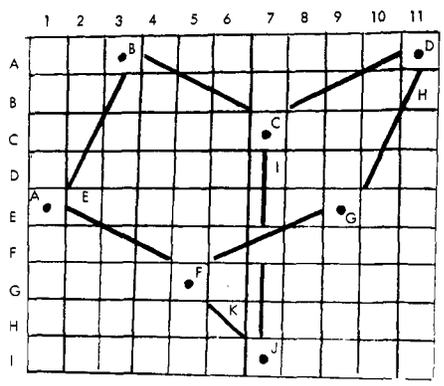



Figure 1.




STRUCTURE CONDENSATION (WITH OCTOBLIQUES) 78 CHARACTERS
(b = BLANK, * = END OF LINE, S = SUPRESS LINE)

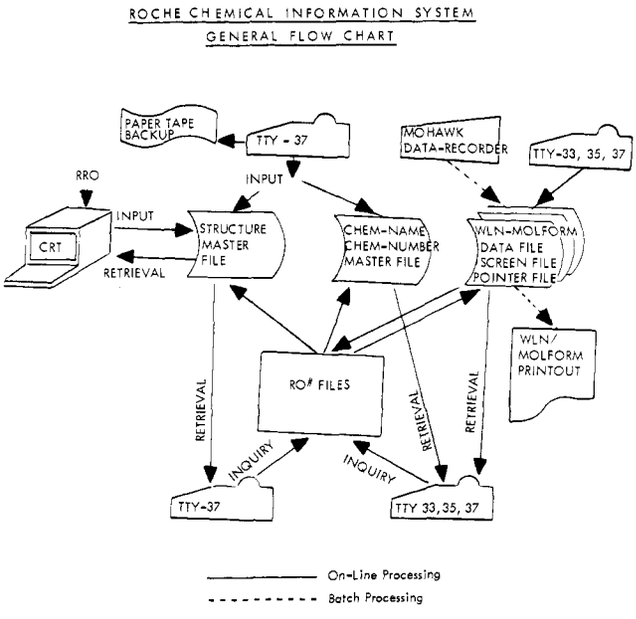



Figure 2.

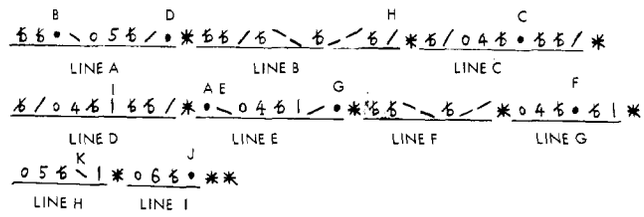



Figure 3.

ring. While the data sheet is being typed, a paper tape simultaneously records the Ro number, the structure, the chemical name, the chemist's name and number, and the molform. The entity of each compound on the paper tape is indicated by a record separator. After the chemist has approved the data sheet, the paper tape is entered on-line through the terminal, and its content is stored on a temporary disk file which is subject to programmed editing routines. These consist of checking the Ro number format and the sequence of the data fields by scanning for key-control words, and of eliminating unnecessary tape codes. Finally, the items of chemical data are released and merged into their respective master files for subsequent on-line search and retrieval.

To store a structure, a special algorithm is employed to translate the random motion which the operator used in the typing process to a line-at-a-time configuration by reconstructing all typing element movements, relative to a starting position, into a two-dimensional array. The information is further condensed by the elimination of three or more consecutive blanks in this matrix. A special code followed by a binary number indicates the number of blank positions to the next nonblank character. All of the right-side blanks of each line are omitted by using an "*" to denote "end of line." This "squeeze technique" shows a 50% gain of storage space (Figure 3). It is calculated that one block of disk space (320 words) stores eight chemical structures or 25 chemical names. When a structure is retrieved and printed on-line on the teletype, the necessary blanks are reintroduced and the original size structure, left justified, is printed line by line.

Empirical formula cards giving both names and structures are prepared by processing the paper tapes in local mode with special perforated card stock in the TTY-37. The empirical formula card catalog is maintained for the convenience of chemists unfamiliar with the Wiswesser line notation.

The Datanet 760 (DTU-760) is the second on-line terminal used in the system. It is a cathode ray tube display plus keyboard device on which a complete "page" is transmitted to and from the chemical information subsystem module. Each page consists of up to 26 lines of data, each line containing a maximum of 46 characters. The keyboard is equipped with special chemical symbols for structure input and retrieval. Figure 4 shows a structure on the CRT screen.

With the implementation of the on-line chemical information system in 1970, the backlog structures were entered through the CRT terminal. Programmed editing was performed on-line so that corrections of Ro number format and sequence could be made immediately. The input routine provided the playback feature so that the structure image just transmitted for processing could be brought back on the screen for redisplay. By retaining the basic diagram and making the necessary modifications on the screen, there was a substantial gain of time and labor when entering a series of similar structures. Using the time-sharing system, two CRTs were operated simultaneously as input terminals when manpower was available. Under ideal environment the average time for entering, editing, and transmitting a structure on the CRT was two minutes. By the end of 1973, RRO had completed entering all the backlog chemical structures and, at present, all input is handled *via* the teletype.

**B. Off-Line Input Device.** The Mohawk Data Sciences

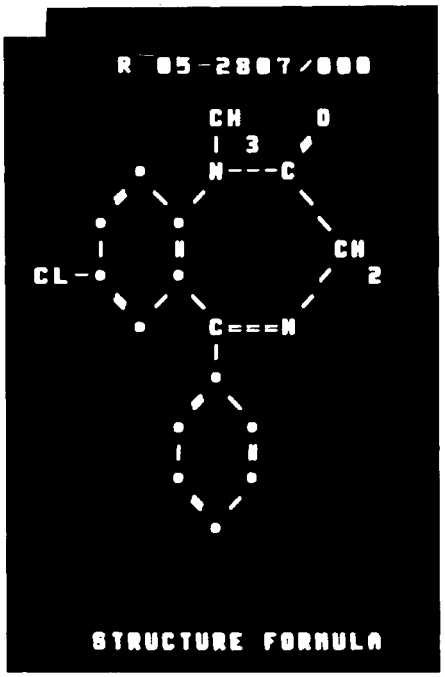



**Figure 4.**

magnetic tape recorder is used for entering the Wiswesser line notation and, optionally, the molecular formula for the registered compounds. During the editing procedure, a "checker program" (a series of subroutines developed by Dow Chemical Company[3] and modified at Roche for in-house usage) performs a syntactical check of each input notation, analyzes and resolves it into its constituent fragments or functional groups, and calculates the molecular formula. The machine-generated molform is compared with the manually calculated molform which has been entered *via* the magnetic or paper tape. When a discrepancy is found, the Ro number, the WLN, and both molforms are printed out for manual evaluation and subsequent correction. The utilization of the WLN file for on-line and batch searching is explained under the "Wiswesser Retrieval System."

## CHEMICAL INFORMATION SYSTEM (CIS) FILE STRUCTURE

The primary control key for accessing all chemical data files is the Ro number with its three component parts: the two-digit prefix, the four-digit basic code, and the three-digit salt code. All chemical data files have identical file format. There is a utilization table at the beginning of each data file and corresponding index file to denote the current status. Since this data base has two attributes (much more information is added than is deleted and search activity is higher than update activity), it is feasible to use direct access files with highly indexed data records for efficient maintenance and search. A minimum of two levels of indexing is provided. The Ro prefix table has 100 entries, each of which stores the address of a base linkage table if that specific prefix is being used; each base linkage table can contain a maximum of 10,000 entries, each pointing to the data location for a compound. If more than one salt code is present, a linked list of tertiary pointers is utilized (Figure 5).

Maintenance of the CIS file is performed, using a direct replacement scheme, by entering the Ro number and the revised data item *via* the terminal. By matching identical Ro number and data field (chemical name, chemical structure, chemist name, etc.), the current entry replaces the information previously stored on the file. Adding new data items to the file is accomplished by updating the pointer and linking the record into its logical sequential position in the file based on its Ro number. A record is deleted from the file by severing its indexing pointer. Since frequent updating generates unnecessary pointers and thus tends to degrade processing, a file utilization report is produced periodically and examined to determine the need to reorganize the data file. Reorganization is accomplished by a straightforward "file-save-and-restore" program using a backup tape file as an interim medium.

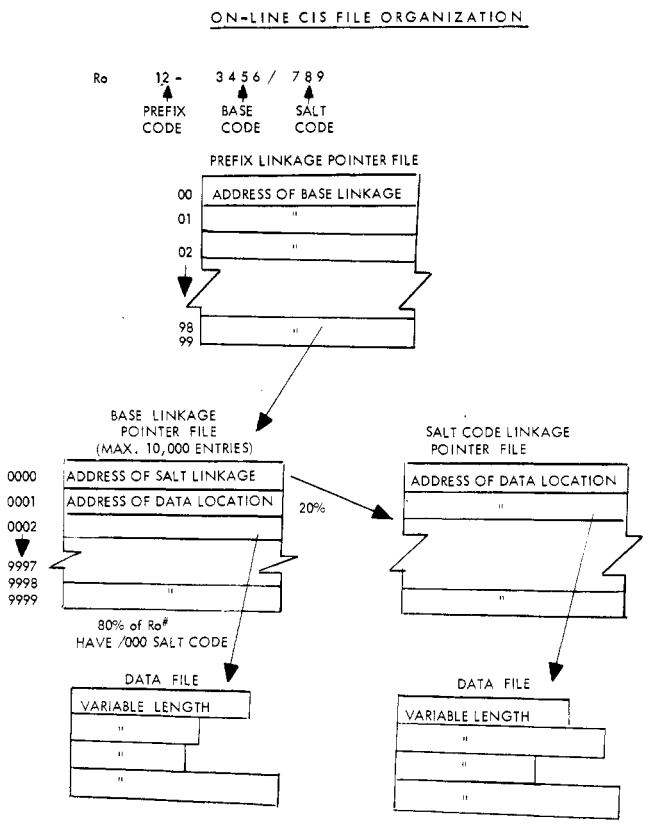


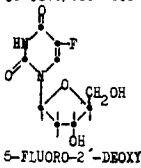

**Figure 5.**

To economize disk space, subset data bases, with the exception of the structure file, are stored in compacted BCD mode. The structure records on the random master file are stored in the modified 128 ASCII code combination with both upper and lower case.

## ON-LINE SEARCH BY Ro NUMBER

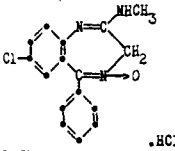
This chemical information subsystem serves primarily as the communication vehicle between various data files. One or more current files can be generated to store Ro numbers entered manually through the teletype or written directly by one of the retrieval programs. The manually entered Ro numbers are internally edited by rejection of input with improper format and addition of the 000 salt code to those numbers entered without it. Other file editing functions which can be carried out on command are additions, deletions, sorting, sorting with removal of duplicates, and copying all or a portion to another file. The Ro numbers serve as the control field for searching and retrieving chemical names, structures, chemist codes and names, Wiswesser notations, and molecular formula. The output of searches is usually presented to the inquirer in the form of lists of structures and/or names printed on the teletype (Figure 6). These lists are often appended to reports by the scientists. The retrieved chemical data can also be displayed on the CRT screen.

05-0360/000 803 DUSCHINSKY,R.,DR.



5-FLUORO-2'-DEOXYURIDINE

05-0690/000 301 STERNBACH,L.H.,DR.



7-CHLORO-2-METHYLAMINO-5-PHENYL-3H-1, 4-BENZODIAZEPINE 4-OXIDE HYDROCHLORIDE

**Figure 6.**

## ABSOLUTE NOTATION

```
ENTER WLN="T6VMVMV FHJ FYIVOI"

# FOUND-        1.

02-0993/000 "T6VMVMV FHJ FYIVOI"

ENTER WLN="T6VMVMV FHJ F7"
NOT FOUND-"T6VMVMV FHJ F7"
```

**Figure 7.**

## LEADING AND TRAILING FRAGMENTS

```
ENTER WLN="T6VMVMV FHJ FYIVOI%

 # FOUND-        3.

02-0993/000 "T6VMVMV FHJ FYIVOI"
02-3625/000 "T6VMVMV FHJ FYIVOI FIQ"
02-0997/000 "T6VMVMV FHJ FYIVOI F2UI"

ENTER WLN=% FYIVOI"

 # FOUND-        2.

02-0993/000 "T6VMVMV FHJ FYIVOI"
02-3625/000 "T6VMVNV FHJ DIQ FYIVOI"
```

**Figure 8.**

## EMBEDDED FRAGMENT

```
ENTER WLN=% FYIVO%

 # OF POSSIBLE RECORDS ARE-     62.

02-0993/000 "T6VMVMV FHJ FYIVOI"
02-3625/000 "T6VMVNV FHJ DIQ FYIVOI"
02-3625/000 "T6VMVMV FHJ FYIVOI FIQ"
02-0997/000 "T6VMVMV FHJ FYIVOI F2UI"
02-0943/000 "T6VMVMV FHJ FYIVO2"
02-0968/000 "T6VMVMV FHJ FYIVO2 F2UI"

 # OF RECORDS ACCEPTED ARE       6.
```

**Figure 9.**

## WISWESSER RETRIEVAL SYSTEM

To a chemist familiar with the WLN, the notation conveys as much information as a two-dimensional structure with the advantage that it can be manipulated and searched directly by computer. At the present, all Roche compounds are coded in Wiswesser notation for registry search, substructure search, and data correlation. Two independent Wiswesser retrieval subsystems are currently in operation within the integrated chemical information system.

### On-Line Wiswesser Retrieval Subsystem

**A. File Organization.** The on-line WLN retrieval system consists of four random access files. Three of these files contain pointers to the WLN data records stored on the fourth file. The three files are fully inverted; that is, each file contains no key information but only the pointers (plus a limited amount of control information defining the size of the file) which impose a desired sequence on the data file, *e.g.*, Ro number, alpha, or permuted[4] sequence. It should be noted that each pointer on the permuted file contains, in addition to the address of the WLN data record, a subpointer to the permuted character within the record. There are 7.2 times as many pointers on the permuted file as are on the Ro number or alpha pointer files. The key information items stored in the WLN data records are the Ro number, the notation, the chemist number, and the notation type classification (used to differentiate tautomers). Searching the file using the Ro number or the alpha sequence notation as the accession key is straightforward. The permuted search is more complex in that the permuted fragments must first be logically shifted using the secondary subpointer in the permuted pointer.

**B. Search Logic.** When searching the WLN file for an Ro number, alpha notation, or permuted fragment, a binary search strategy is used. In order to search a file, the first and last records are examined. If the desired record is located between them, then the record midway between them is examined to determine if the record being sought is above or below the middle record. After this determination has been made, a new range is selected which is half the size of the old range and the test is repeated. This method is continued until the desired record is found or all records on the file are exhausted. A binary search strategy allows for the rapid searching of very large data bases. As the size of the data file doubles, the search time will increase by only one additional record access (two reads to mass storage).

Four options are available for notation searching as illustrated in Figures 7, 8, and 9.

| | |
|---|---|
| "XXXXXX" | absolute notation |
| "XXXXXX% | leading fragment |
| %XXXXXX" | trailing fragment |
| %XXXXXX% | embedded fragment. |

A registry search using the absolute notation option is made before assignment of an Ro number to a compound. Fragment searches are used when a quick count is needed on the frequency of occurrence of a particular string of characters to plan search strategy or when the inquiry would be satisfied by a list of all compounds containing the fragment.

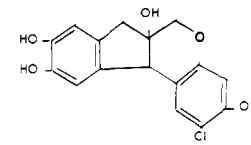
Generic searches applying "and" and "or" logic to a series of notation fragments can be performed using the WLN alpha and permuted files. These operations are accomplished by first collecting a list of pointers for each desired operand (*i.e.*, notation fragment) and then logically operating on these lists (Figure 10). The pertinent Ro numbers and notations can be printed on the teletype and/or the Ro numbers can be stored on a a disk file for subsequent application. The method is effective in generic searches requiring a series of unvarying character strings as for cyclic compounds but less effective in those satisfied by disparate notations as for varyingly substituted aliphatic or benzene compounds. An advantage of this on-line search

LOGIC SEARCH

```
WLN LOGIC SYSTEM
ENTER WLN=% T6VM%
ENTER WLN=% T6VN%
ENTER WLN=% FYIVO%

ENTER LOGIC OPERATIONS
=W01+W02
=L01*W03

6 RECORDS HAVE BEEN FOUND FOR LOGIC OPERAND #  2

02-0943/000 "T6VMVMV  FHJ  FYIV02"
02-0968/000 "T6VMVMV  FHJ  FYIV02 F2UI"
02-0993/000 "T6VMVMV  FHJ  FYIVOI"
02-0997/000 "T6VMVMV  FHJ  FYIVOI F2UI"
02-3625/000 "T6VMVNV  FHJ  DIQ FYIVOI"
02-3625/000 "T6VMVMV  FHJ  FYIVOI FIQ"
```

**Figure 10.**

BIT SCREEN PRESENTATION

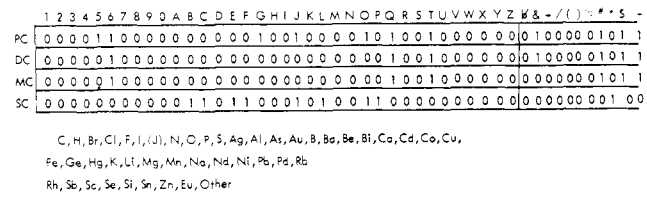



WISWESSER NOTATION

T C6 B566 LO&TT&J EQ FQ JQ OQ PG

MOLECULAR FORMULA C16 H13 O5 Cl

FIXED POSITION BIT SCREEN



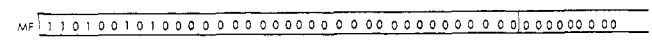




**Figure 11.**

technique is the speed with which it provides an answer. For searches involving lists of only a few thousand pointers, the retrieval time, depending on the status of the time-sharing system, ranges from a few seconds to a few minutes for the file of over 100,000 data records. However, there are two limitations to this retrieval system: (1) each operand must contain at least one permuted symbol; (2) if the first symbol in the operand is not permutable, or if one or more of the retrieved lists is high in volume (over 10,000 records), search time can become excessive.

## On-Line/Batch Wiswesser-Molform Bit Screen Retrieval Subsystem

**A. System Concept.** The original concept, proposed by W. J. Wiswesser and developed by ISI (Institute for Scientific Information) of Philadelphia, is based on the notion of generating and comparing bit screens (fixed-position bit marks) for the search profiles and the WLN data records.[5] This presearch by bit screen permits bypassing most of the WLN records and selecting only a small number of relevant records for detailed examination. The technique has been found to be highly effective in reducing computer processing time for substructure searches. Based on in-house needs, a series of programs was developed to interface on-line with batch processing of the WLN data base. The search queries are entered on-line *via* the teletype terminal and the compacted Wiswesser data records and their corresponding indices (pointers) are generated by the software package "Indexed Sequential Processor" which permits both sequential and random accessing of the files within the same program.

**B. Bit Screen Generation.** Five different sets of fixed-position bit marks are generated from each record on the data base and from each profile entered by the user for the presearch activity. Four sets of screens are based on the occurrence of WLN symbols while the fifth set arises from the occurrence of atoms in the molecular formula as illustrated in Figure 11.

- PC Primary Character. The WLN symbol appears at least once within the notation.
- DC Double Character. The WLN symbol appears at least twice within the notation.
- MC Multiple Character. The WLN symbol appears more than twice within the notation.
- SC Space Character. The WLN symbol is preceded by a space within the notation (*e.g.*, a locant).
- MF Molform Screen. The atomic symbol appears in the molecular formula of the compound.

**C. Profile Construction.** A profile is constructed by first listing a series of terms which can be atomic symbols or WLN fragments. The atomic symbol terms include the relational expressions "less than," "equal to," or "greater than." The Boolean operators "or," "and," "not," "followed by," or "same word" are used with the WLN fragments. To increase search capabilities the following special characters can also be used

- # representing any numerical symbol 0–9
- + denoting any alpha character A–Z
- $ indicating either alpha or numerical symbol A–Z or 0–9
- * connoting any locant

Each of the terms is assigned a weight (equal for "or"; different for "and") and the profile is assigned a threshold weight such that, for a record containing the desired combination of WLN fragments and elements, the sum of term weights will equal or exceed the threshold weight. Complex searches can be entered as a series of subprofiles which use the same threshold weight and the same or different screens.

As the second step in the construction of a profile, screens are chosen by citing any atomic symbols which will significantly reduce the number of potential hits and by listing primary, double, multiple, and space characters appearing or implied in the WLN fragments. If the search is to be restricted to a certain range or set of Ro numbers, the qualifying limits are specified.

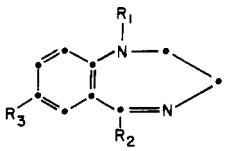
After a profile is entered *via* the teletype in free format, editing routines check for syntactical errors such as missing term weight, unachievable threshold weight, etc. (Figures 12 and 13). When all profiles, a maximum of twenty, have been entered, a 24-hour turnaround time is allowed for the batch processing of the search.

**D. Search.** Searching the WLN data base is done in two steps. The first step compares the bit screen files of the data base and the profile. When the screen records are identical, the data record undergoes any Ro qualifier test specified in the profile to further eliminate irrelevant records and reduce the number of potential hit records.

The second step is the conventional string character search which examines the potential hit records in different combinations of logical or syntactical relations as indicated

PROFILE CONSTRUCTION

QUERY I



$R_1$ = alkyl-N (tert. amine)

$R_2$ = fluorophenyl

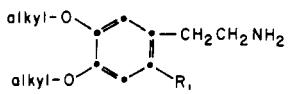
$R_3$ = Cl or Br

```
(100)(WLN)T67 GN(FB) JN(SW)J(40)
 CG(25) CE(25) G#N(20)
 KR *F(15)
(PC=T67JRF)(MC=N)(SC=CGJK)

Qualifying Ro prefixes 5 & 20 only
```

**Figure 12.**

PROFILE CONSTRUCTION

QUERY II



$R_1$ = amino
hydroxy
halogen (chloro, bromo)
alkyl = Me, Et, nPr

```
(130)(WLN)Z2R(40)
 BZ(25) BQ(25) BG(25) BE(25)
 DO#(20) EO#(15)
(NOT) C(1)(NOT) F(2)
(MF)C<15(30)
(PC=R2Z)(DC=O)(SC=BDE)
```

**Figure 13.**

by the search profile. For each matching term, the corresponding term weight is added to an accumulator; in the case of a "not" term, the weight is subtracted. If the final achieved weight equals or exceeds the user-designated threshold weight, the data record is selected as an "actual hit" and is written on the output retrieved file. The pertinent WLN/Molform records for each profile are sorted into Ro number sequence. If desired, the records can be sorted into achieved weight sequence. This is useful when each subprofile has been designed so that its achievable weight is a reflection of the pertinence of its hits to the inquiry.

Our experience shows that the bit screen file comparisons are made at the rate of about 200,000/min and the string character searches at about 2000/min.

The retrieved WLN/Molform data are printed in batch environment on the printer as shown in Figures 14, 15, and 16. The Ro numbers of the retrieved records are retained on a permanent disk file, and, after the relevance of the retrieved records has been evaluated by inspection, the file can be edited and then used to retrieve the requested chemical and biological data.

## CONCLUSION

Operation of the on-line chemical information system affords a convenient means to provide chemists and biologists with printed lists of chemical structures and/or

```
PROJECT R1044  WLN-MOLFORM RETRIEVAL SYSTEM

PROFILE-1RCR    (100) (WLN)T67 GN(FB) JN(SW)J(40) CG(25) CE(25) G#N(20)
                KR *F(15) (PC=T67JRF) (MC=N) (SC=CGJK)

05-6603/000  100  A  303  T67 GNV JN&TJ CG G2N2&2 J KR BF     C22H27N3OFCL
20-6271/000  100  A  303  T67 GN JNJ CG G2N2&2 HS1 KR BF     C22H25CLFN3S

                POTENTIAL HITS PER PROFILE    37
                SEARCH HITS PER PROFILE        2
                END OF PROFILE-1RCR
```

**Figure 14.**

```
PROJECT R1044  WLN-MOLFORM RETRIEVAL SYSTEM

PROFILE-2RCR    (130) (WLN)Z2R(40) BZ(25) BQ(25) BG(25) BE(25) DO#(20) EO#(15)
                (NOT) C(1)(NOT) F(2) (MF) C<15(30) (PC=R2Z) (DC=O) (SC=BDE)

05-9538/000  130  A  408  Z2R  BE  DO1  EO1          C10H14NO2BR
05-9913/000  130  A  401  Z2R  BG  DO1  EO1          C10H14NO2CL
06-2613/000  130  A  153  Z2R  BZ  DO1  EO1          C10H16N2O2

                POTENTIAL HITS PER PROFILE   124
                SEARCH HITS PER PROFILE        3
                END OF PROFILE-2RCR
```

**Figure 15.**

```
PROJECT R1044  WLN-MOLFORM RETRIEVAL SYSTEM

                TOTAL INPUT RECORDS          9
                TOTAL PROFILES               2
                  PROFILE-1RCR   PREFIX   QUALIFIER 05 20
                POTENTIAL HITS TALLY RCDS    2
                TOTAL SEARCH HITS            5
```

| PROFILE-NO | PROFILE-ID | PRE-QUAL. HITS | POTENTIAL HITS | FALSE HITS | ACTUAL HITS |
|---|---|---|---|---|---|
| 1 | 1RCR | 99 | 37 | 35 | 2 |
| 2 | 2RCR | 124 | 124 | 121 | 3 |
| | TOTAL | 223 | 161 | 156 | 5 |

**Figure 16.**

names. It also allows inclusion of structures in the empirical formula card catalog.

Using the on-line Wiswesser notation subsystem, registry searches are done with speed and accuracy, independent of the variations of nomenclature. Generic searches satisfied by a constant Wiswesser character string are made on-line with the advantage of an immediate answer.

When more complex searching is required or when the volume of pertinent data is great, the on-line/batch Wiswesser molform subsystem provides an economical method of searching a large file by application of the bit screen technique.

The development of the chemical information system is a continuing process. Improvements in hardware, software, and information searching techniques are utilized to meet the ever changing requirements of the scientists and to achieve maximum efficiency in handling the rapidly increasing volume of data.

## ACKNOWLEDGMENTS

We wish to acknowledge the valuable programming contributions of Mr. J. H. Mianecki. The encouragement and direction of Dr. E. C. Foernzler, Dr. W. R. Sullivan, Mr. J. J. Smith, and Mr. R. L. Zachmann toward the successful implementation of the chemical information system is appreciated. In addition, we are very grateful for the expert advice we received from Dr. P. F. Sorter who has had many years of broad experience with the Wiswesser line notation.

## LITERATURE CITED

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

(2) Gottardi, R., "A Modified Dot-Bond Structural Formula Font with Improved

Stereochemical Notation Abilities," *J. Chem. Doc.*, **10**, 75–81 (1970).
(3) Bowman, C. M., Landee, F. A., and Reslock, M. H., "A Chemically Oriented Information Storage and Retrieval System. I. Storage and Verification of Structural Information," *J. Chem. Doc.*, **7**, 43–47 (1967).
(4) Sorter, P. F., Granito, C. E., Gilmer, J. C., Gelberg, A., and Metcalf, E. A., "Rapid Structure Searches Via Permuted Chemical Line-Notations," *J. Chem. Doc.*, **4**, 56–60 (1964).
(5) Granito, C. E., Becker, G. T., Roberts, S., Wiswesser, W. J., and Windlinx, K. J., "Computer-Generated Substructure Codes (Bit Screens)," *J. Chem. Doc.*, **11**, 106–110 (1971).

# Automated Conversion of Chemical Substance Names to Atom-Bond Connection Tables

G. G. VANDER STOUW,* P. M. ELLIOTT, and A. C. ISENBERG

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received August 1, 1974

**Chemical Abstracts Service (CAS) has developed a computer program for converting systematic names of organic compounds into atom-bond connection tables of the type input to the CAS Chemical Registry System. This program, called the nomenclature translation program, is designed to process names which are based on the word roots and punctuation conventions used in CA Index nomenclature. Both inverted and uninverted names can be processed if they are specific and unambiguous. Inconsistent or ambiguous names will be rejected, with appropriate diagnostics, as will names containing features not recognized by the program. The translation program is currently being installed as part of a comprehensive name editing system.**

Chemical substances may be described in several ways, including both structural diagrams and a variety of chemical names. The names used for a given substance may include both "systematic" names, which are constructed from commonly understood nomenclature fragments that correspond to fragments of the structural diagram, and also other names which do not describe the structure of the substance to which they refer. For example, the substance described by the simple structural diagram $N{\equiv}C{-}CH_2{-}CH_2{-}C{\equiv}N$ has been referred to not only by structurally descriptive names including "butanedinitrile", "succinonitrile", and "1,2-dicyanoethane", but also by trade names such as "Dinile", "Deprelin", and "Suxil".

The Chemical Substance Index to *Chemical Abstracts* (CA) brings together, under a single name, all CA references to a particular chemical substance which has been selected as a CA Index entry regardless of the various names used in the original documents. The substance appears in the Index at the CA Index Name, a "canonical" name derived by the application of a rigorous and comprehensive set of systematic name selection rules. Preparation of CA Indexes is supported by the Chemical Registry System, a computer-based system which links the structure and various names of a substance; this system included approximately 2.7 million substances at the beginning of 1974. The address of a substance in this System is the CAS Registry Number, a unique identifying number which is associated with a canonically numbered atom-bond connection table in the system's structure file[1,2] and with the CA Index Name and other names for that substance in the Registry nomenclature file.[3]

There are two basic routes for retrieval of substance information from the CAS Registry files (see Figure 1). One is "name match", in which a name is compared against the contents of the name file; if a "match" occurs, the Registry Number is retrieved. The other basic route is structure registration, in which a keyboarded structural diagram is converted to the canonical connection table which is matched against the structure file to retrieve the Registry Number. If no "match" occurs during structure registration, *i.e.*, the substance is new to the file, a new Registry Number is automatically assigned. The retrieved Registry Numbers can then be used to retrieve the CA Index Name and any other names from the Registry name file.

Although the Registry System links the names and structural representations of a substance, name and structure have not previously been directly interconvertible. We report here the development of a computer program for converting chemical names into connection tables, a process we call "nomenclature translation".† As illustrated by the dotted lines in Figure 1, this process provides an alternate method of structure registration by allowing a new substance to be input *via* a structurally descriptive systematic name instead of only as a connection table taken from a structural diagram.

There are two major potential applications for the use of nomenclature translation as an entry to Registry processing. One is for entering new substances for which systematic names are available into the CAS Registry structure file, bypassing the need for input of structural diagrams. The other application is to verify that the structural records and the CA Index Name on file for a given substance are fully consistent. Processing the CA Index Name for a substance by nomenclature translation, followed by registration of the resulting connection table, should lead to retrieval of the Registry Number previously assigned to that substance. If the expected number is not retrieved, an inconsistency between the name and structure records on file is indicated and the records in error must be identified and corrected. Nomenclature translation thus can provide a powerful tool for use in a system for editing CA Index

* To whom correspondence should be addressed.

† The conversion of connection tables to names, called "nomenclature generation", has been described for bridged ring systems[4,5] and is the subject of continuing CAS investigation.