

- (3) Bowman, C. M., F. A. Landee, M. H. Reslock, and B. P. Smith, "Automatic Generation of Structural Fragment Codes From the Wiswesser Line Notation for Rapid Structure Searches," Proceedings of the Wiswesser Line Notation Meeting of the Army Chemical Information and Data Systems, James P. Mitchell, Ed., pp. 49-56, EASP 400-8, Edgewood Arsenal, Md., 1968.
- (4) Farris, R. N. "Computers Cut the Cost of Literature Searches," *Chem. Eng. Progr.* **62** (5), 89-91 (1963).
- (5) Hyde, E., F. W. Matthews, L. H. Thomson, and W. J. Wiswesser, "Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds," *J. CHEM. DOC.* **7**, 200-204 (1967).
- (6) Landee, Franc A., "Computer Methods of Handling Files of Chemically Oriented Information," unpublished paper presented in Moscow, USSR, Oct. 1965.
- (7) Landee, Franc A., "Computer Programs for Handling Chemical Structures Expressed in Wiswesser Notation," Presented before the Division of Chemical Literature, 147th Meeting ACS, April 8, 1964.
- (8) Opler, A., and T. R. Norton, "A Manual for Programming Computers for Use with a Mechanized System for Searching Organic Compounds," The Dow Chemical Co., Western Division, Pittsburgh, Calif., 1956.
- (9) Smith, E. G., *The Wiswesser Line-Formula Chemical Notation*, McGraw-Hill, New York, 1968.
- (10) Smith, E. G., "Machine Sorting for Chemical Structures," *Science* **131**, 142-146 (1960).

## *Index Chemicus Registry System: Pragmatic Approach to Substructure Chemical Retrieval\**

EUGENE GARFIELD, GABRIELLE S. REVESZ, CHARLES E. GRANITO,  
HAYES A. DORR, MARIA M. CALDERON, and ANDREA WARNER  
Institute for Scientific Information (ISI), Philadelphia, Pa. 19106

Received July 21, 1969

**The *Index Chemicus Registry System (ICRS)*, launched in 1968 with the support of a dozen industrial and government organizations, is now a current operational monthly service. Subscribers receive magnetic tapes and printouts, in which the weekly issues of *Index Chemicus (IC)* have been encoded in Wiswesser Line Notations (WLN). Over 13,000 compounds per month are provided in machine language. The canonical WLN is also provided in alphabetized printouts. Encoding of over 400,000 new chemical compounds from *IC* has already been completed, including all those reported in 1967, 1968, and 1969. Since the tapes also include title and other bibliographic information, this paper describes the use of supporting software provided for SDI search systems employing "word" and other searching terms, in addition to the WLN fragments. Use of the monthly and annual printouts are illustrated for those searches which do not require computer manipulation.**

The *ICRS* is designed to provide chemists with current and retrospective chemical information reported in the *IC*.

As *IC* has been described elsewhere,<sup>1</sup> it is sufficient to state that *IC* provides detailed abstracts of journal articles which report new chemical compounds or new chemical reactions.

The *ICRS* has, as yet, not been described in the literature, and a brief description of its main characteristics is necessary to enable one to understand how to search for substructures, both currently and retrospectively.

*ICRS* consists essentially of four data files: WLN magnetic tapes, *IC* bibliographic tapes, WLN printouts, and *IC* weekly issues.

The WLN magnetic tapes contain unique WLN structural descriptions of all new compounds reported in the *Index Chemicus* and are arranged in abstract number sequence. The WLN tapes also contain molecular formulas and *IC* registry numbers, which identify a specific line

in the numbered *IC* abstract where a structural diagram and other information is given.

The *IC* bibliographic tape provides, in machine language, most of the information provided in the printed *IC*: bibliographic citations; codes for new reactions and analytical instrumentation; subject-index terms which are assigned by chemists and include terms related to the properties, uses, and biological activity of the compounds.

The WLN printout version is alphabetized according to the WLN, to provide easy scanning for similar type compounds. The corresponding article from the *IC* can be identified through the registry numbers associated with the notation.

Many searches can be done by simply referring to the monthly or annual *ICRS* printouts. The search for substituted adamantanes is shown in Figure 1. The WLN notation for adamantanes is L66 B6, etc. The *ICRS* printout identifies abstract number 101318, which contains several adamantanes, each of which is separately encoded. The *IC* abstract is shown in the lower portion of Figure 1.

The printouts are also used to formulate machine-search questions.

\*Presented in part at the ACS MARM Meeting, Washington, D. C., February 14, 1969, and before the Division of Chemical Literature, 157th Meeting, ACS, Minneapolis, Minn., April 16, 1969.

## PRAGMATIC APPROACH TO SUBSTRUCTURE CHEMICAL RETRIEVAL

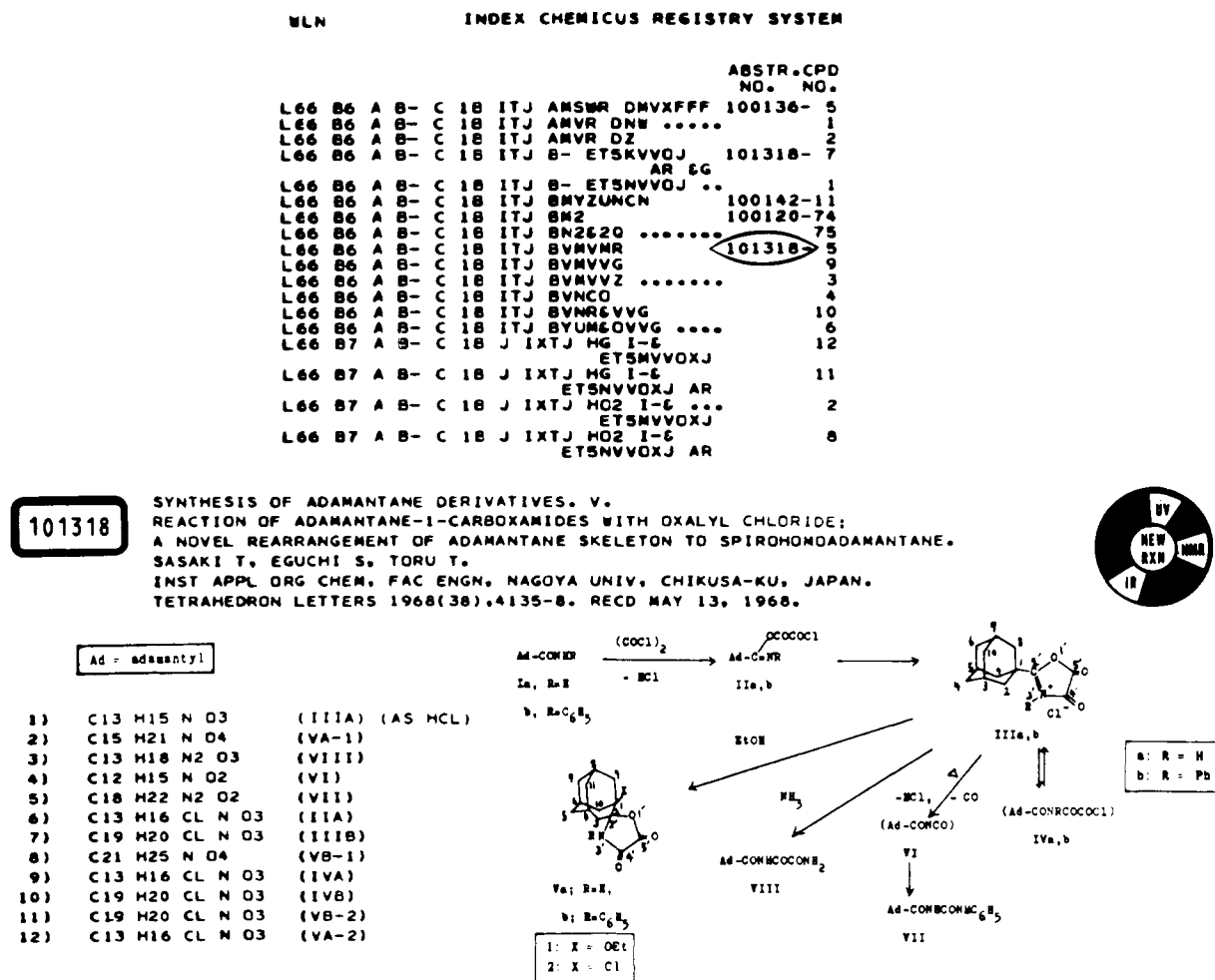


Figure 1. ICRS printout showing substituted adamantanes and corresponding IC abstract

The WLN has been extensively described elsewhere.<sup>2</sup> The WLN tapes can be searched for both specific and generic structures. While there are many instances, especially in SDI systems, in which the *IC* bibliographic tapes are used to augment substructure searches, the primary substructure searching capability is derived from the WLN tapes. For example, a generic search could be conducted for all ortho-substituted aniline compounds. Figure 2 shows several such compounds, their WLN codes, and *ICRS* WLN listings. In the monthly *ICRS* printout, all ring systems contained in a compound are listed as separate entries. Since many searches include a particular ring system as a starting point, the printouts allow for a considerable amount of manual searching. Generating additional entries for each ring system produces a printout which contains approximately 8.5% more notations than the number of compounds on the WLN magnetic tape. In Figure 3, the two separate notations for the WLN's are shown for a compound containing two ring systems.

The *IC* bibliographic tape can be used to further qualify the results of a WLN tape search. If a search of the WLN file produces a large number of candidate compounds, the *IC* bibliographic file can be used to indicate those papers in which certain biological or other properties, activities, uses, or analytical methods have been reported.

The *IC* weekly issues can then be used for making the final selections from the candidates produced by the computer search of tapes or manual searches of the printouts. The most important selection criterion an *IC* printed abstract provides is the structural diagram of the compound or the flow diagram of the reaction in which the compound occurs. *IC*'s abstract usually also contains an author-prepared summary of the paper.

## THE WISWESSER LINE-FORMULA NOTATION (WLN)

To fully utilize the substructure search capabilities of *ICRS*, one must understand the Wiswesser Line Notation. However, it is not necessary for a chemist to master every detail of WLN in order to use the results of a search, although an understanding of the symbols and the principles of the notation is helpful. Furthermore, the liberal use of structural diagrams in the monthly printouts facilitates searching.

Figure 4 shows the lists of WLN symbols which appears in each issue of *ICRS*.

In Figure 5 some 1,2-substituted 3-pyrazolidones are shown. These are found in the printed version of the WLN tape under T5NNVTJ. This section of the printout is easily accessible, once the term pyrazolidones is included

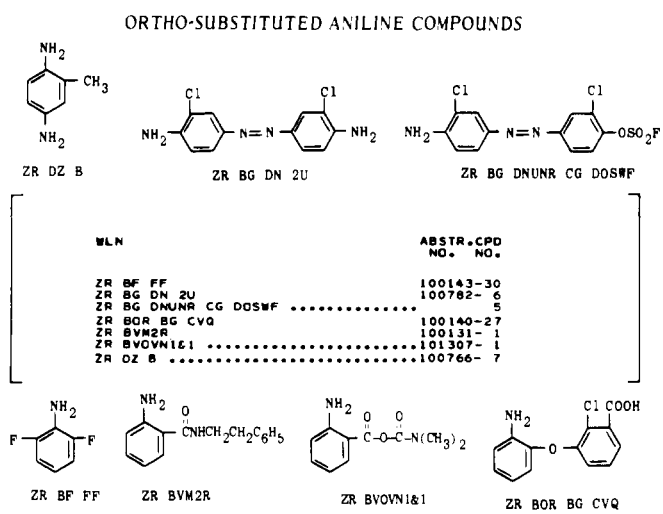
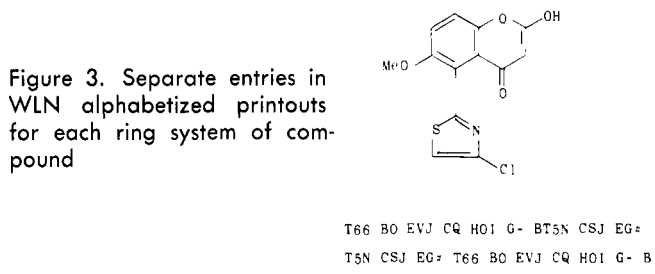


Figure 2. Structural diagrams for seven compounds, together with their WLN notations and the section of ICRS showing the IC registry numbers



in a search dictionary. Such a dictionary for many common generic structures with their appropriate notations has been incorporated in the ICRS printed issues for this very purpose (Figure 6). Once the parent structure is found, it is relatively easy to visually scan the remainder of each notation for substituents.

It is one thing to provide a machine-reliable file of encoded compounds and subject descriptions—it is another thing to use that file to provide information to the ultimate user. To use such a file on a computer obviously requires programs or software—and this implies a system. The RADIICAL<sup>3</sup> System is the software component of the ICRS.

RETRIEVAL AND AUTOMATIC DISSEMINATION OF INFORMATION FROM INDEX CHEMICUS AND LINE NOTATIONS (RADIICAL)

There have been many programs written for machine searching of a WLN file. For example, Hyde<sup>4</sup> has described programs to convert line notations to a connectivity matrix. Fraction<sup>5</sup> and Kulpinski<sup>6</sup> have also written algorithms which convert WLN to connectivity tables. Finlay<sup>7</sup> developed programs for a "Permuted WLN" Index, as did Granito *et al.*<sup>8</sup>

We have developed a software system which can be used in two modes—one is for ISI source tapes, a by-product of the *Science Citation Index* (SCI). We call that mode ISIS, an acronym for ISI Search. The other mode is for ICRS tapes. We call that mode RADIICAL—

- A alkyl
- E bromine
- G chlorine
- J halogen
- K quaternary nitrogen
- L carbocyclic ring
- N secondary nitrogen
- Q tertiary nitrogen atom
- R hydroxyl
- 4 benzene ring
- T heterocyclic ring
- U double bond
- V carbonyl
- W nonlinear dioxo
- X quaternary carbon
- Y tertiary carbon
- Z amino or amido
- & punctuation mark
- separator or connective
- / multiplier stop

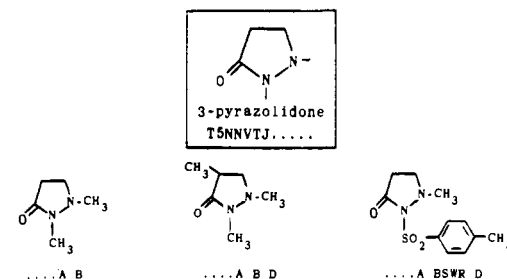
Numerals preceded by a space are multipliers of preceding notation symbols; or within ring signs L...J or T...J show the number of multicyclic points in the ring structure.

Numerals not preceded by a space show ring sizes if within the ring signs; elsewhere numerals show the length of internally saturated, unbranched alkyl chains and segments.

Letters following a space and hyphen are proposed as symbols with special meanings to denote stereoisomerism.

All international atomic symbols except K, U, V, W, Y, Cl, and Br are used.

Figure 4. WLN symbols



INDEX CHEMICUS REGISTRY SYSTEM

WLN	ABSTR. CPD NO.	M.F. NO.
T5NNVTJ A B	99543-9	10
T5NNVTJ A B D	10	17
T5NNVTJ A B E	99620-3	2
T5NNVTJ ASUR D& B	99543-15	16
T5NNVTJ AY BY	20	11
T5NNVTJ AY BY D	12	13
T5NNVTJ AY BY E	18	14
T5NNVTJ A2 B2	13	14
T5NNVTJ A2 B2 D	19	19
T5NNVTJ A2 B2 E		
T5NNVTJ A3 B3		
T5NNVTJ A3 B3 D		
T5NNVTJ A3 B3 E		

Figure 5. Substituted 3-pyrazolidones

COMPOUND	WLN	PAGE	COMPOUND	WLN	PAGE
Acenaphthene	L566 1A LT&J	12	Fluorene	L 8656 H+	4
Acridine	T C666 BNJ	27	Furan	T50J	39
Adenine	T56 BM DN FN HNJ IZ	44	Guanine	T56 BM DN FMYMVJ GUM	44
Adenosine	T56 BN DN FN HNJ IZ	45	Hydantoin	T5MVVJ EHV	33
Adrenosterone	L E5 8666 CV FV QV	5	Imidazole	T5M CNJ	33
Androstane	MUTJ A E	7	Indan	L58T&J	12
Aniline	ZR	66	Indazole	T56 BMNJ	44
Anthracene	L C866J	5	Indole	L56 BHJ	12
Antraquinone	L C666 BV IVJ	5	Indole	T56 BMJ	44
Anthrone	L C666 BV IHJ	5	Isoquinoline	T56 CNJ	62
Azulene	L57J	13	Maleic Anhydride	T5VOVJ	42
Benzaldehyde	VHR	64	Morpholine	T6M DOTJ	49
Benzamide	ZVR	66	Naphthalene	L E5 C666J	8
Benzanthrone	L C666 1A Q IVJ	28	Naphthalene	L6J	18
Benzene	ER	1	Phenanthrene	L 8666J	4
Bromo	QVR	21	Phenanthroline	T C666 CN NJJ	25
Carboxy	GR	2	Phenazine	T C666 BN INJ	27
Chloro	NCR	19	Phenothiazine	T C666 BM ISJ	27
Cyano	QR	21	Phenoxazine	T C666 BM IOJ	27
Hydroxy	QIR	22	Phthalazine	T66 CNNJ	62
methyl	IR	4	Phthalimide	T56 BVMVJ	48
Iodo	IR	68	Piperazine	T6M DMJ	49
Methyl	1OR	67	Piperidine	T6MTJ	49
Methoxy	WNR	64	Pipecoline	T66 BN DN GN NJJ	59
Nitro	T56 BM DNJ	44	Purine	T56 BM DN FN HNJ	59
Benzimidazole	T56 BOJ	47	Pyrazine	T6N DNJ	51
Benzofuran	T66 BO CHJ	60	Pyrazole	T5MNJ	49
Benzopyran	T56 BM DSJ	46	Pyrene	L666 B6 2AB PJ	19
Benzothiazole	T56 BMNNJ	44	Pyridazine	T6NNJ	52
Benzotriazole	L C6 8666J	5	Pyridine	T6NJ	51
Benzphenanthrene	TSOVTJ	41	Pyrimidine	T6N CNJ	50
Butyrolactone	T7MVTJ	63	Pyrolo	T5MJ	33
Caprolactam	T 8656 HMJ	25	Pyrolo	T5MTJ	33
Carbazole	L E5 8666TJ A E FY&ZY	60	Pyrolo	T5M CUTJ	33
Cholestane	T56 BNNJ	47	Pyrolo	T66 BN DNJ	59
Cinnoline	T56 BOT&J	47	Quinoxaline	T66 BNJ	60
Coumaran	T66 BOVJ	47	Quinone	L6V DVJ	15
Coumarin	L6TJ	13	Quinoxaline	T66 BN ENJ	59
Cyclohexane	L6VTJ	15	Succinic Anhydride	T5VOVTJ	42
Cyclohexanone	L6UTJ	15	Succinimide	T5VMVTJ	42
Cyclohexene	L5 AHJ	10	Sulfolane	T5SWTJ	42
Cyclopentadiene	L6VTJ	15	Tetrahydrofuran	T50TJ	39
Cyclopentane	L6UTJ	15	Tetrahydropyran	T60TJ	55
Cyclopropane	L5 TJ	10	Tertralin	L6&TJ	18
Decalin	L66TJ	19	Thiamorpholine	T6M DSJTJ	49
Dioxane	T60 DOTJ	55	Thianaphthene	T56 BSJ	47
Estrone	L E5 8666 FVTTT&J E OO	5	Thiazole	T5N CSJ	34
Ferrocene	L50J O.FE-OL50J	11	Thiophene	T5SJ	42
			Triphenylene	L 86 H666J	4
			Xanthine	T56 BM DN FMYMVJ	44
			Xanthone	T C666 BO IVJ	27

Figure 6. Dictionary of parent structural units frequently encountered and corresponding WLN fragments taken from ICRS for May 1969

## PRAGMATIC APPROACH TO SUBSTRUCTURE CHEMICAL RETRIEVAL

an acronym for Retrieval and Dissemination of Information by *IC* and Line-notations. ISIS and RADIICAL are completely compatible. The basic software is written for IBM 360/30, but is easily modified to a larger 360 system. RADIICAL uses Assembly Language Programming (ALP) and requires a 32K memory. The RADIICAL documentation and flow charts can be used to write programs for other computers.

Formulation of the interest profiles for the RADIICAL system is quite flexible. Search terms may be names of authors, journal titles (or portions thereof), subject indexing terms, words in article titles, portions of words or stems, word combinations, organizational or industrial names, etc. Search elements may be weighted, or combined in a multitude of logical (combinational) or syntactical relationships. Thus, one can specify the order in which certain search elements must occur, and the environment in which they must occur.

RADIICAL software consists of the following programs shown in the flow chart (Figure 7).

1. The question (profile) update program can delete, add, or change profile questions on the input tape. The resulting output tape is acceptable to the search program. It is in account profile number sequence.
2. The search program can search for:
  - (a) WORDS (one or more characters preceded and followed by a blank space).
  - (b) STRINGS (one or more characters, including spaces, in sequence). Strings include word phrases, stems, and segments of the notation.
  - (c) PREFIXES (initial stem), meaning a series of characters in sequence, but preceded by a space.
  - (d) STEMS, a series of characters not containing a space.

The type of logic used with the RADIICAL search program can include:

- (a) AND logic;
- (b) OR logic (including exclusive or);
- (c) NOT logic (indicating absolute NOT or weighted NOT);
- (d) SAME WORD logic (looking for two search terms, but only if in the same word);

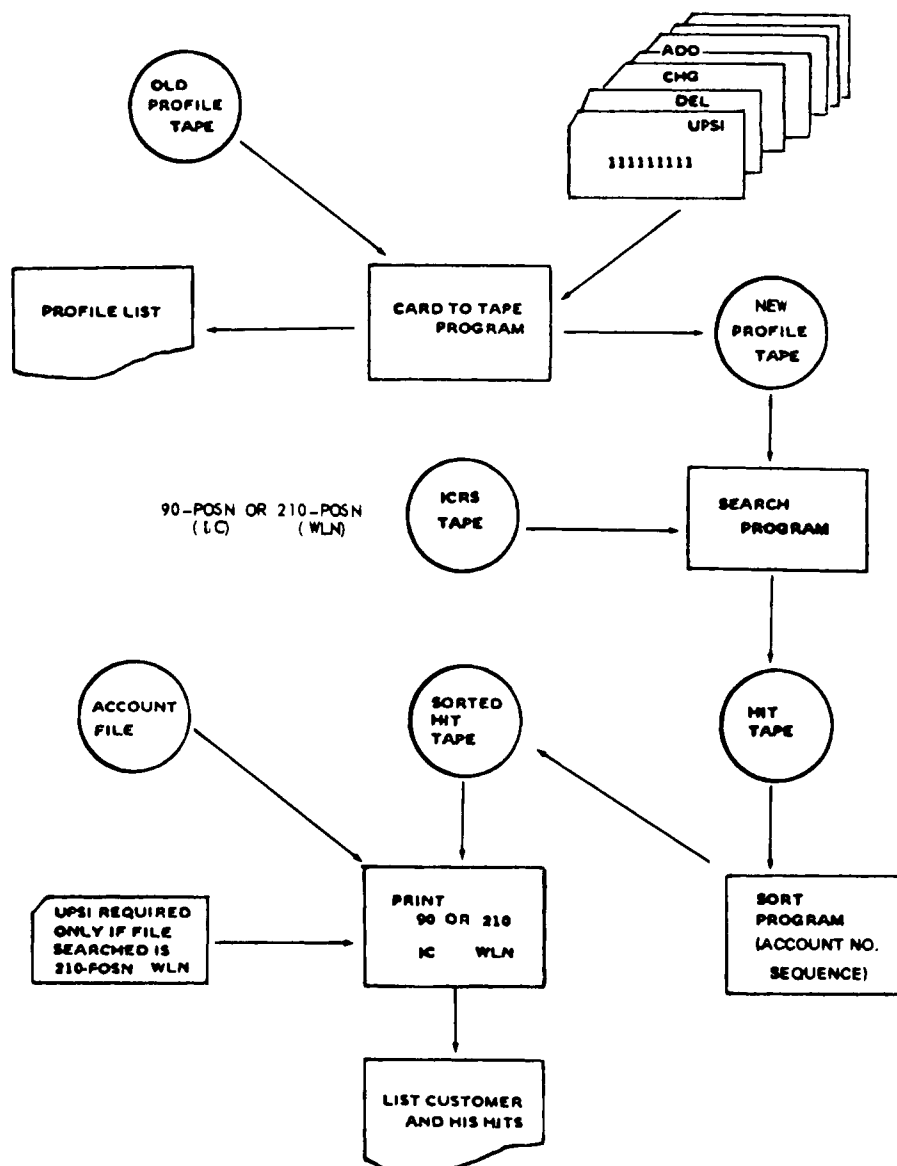


Figure 7. ICRS RADIICAL software system flow chart

- (e) FOLLOWED BY logic (two words, strings, stems, or prefixes, but only if one follows the other, rather than precedes)

The search program results in a "Hit" tape, which must be sorted under control of the sort control card deck.

3. The sort control program and card deck sorts the "Hit" tape. The sort run arranges hits in order by account number. The sorted hit tape then is used with the printout program.

4. The printout program prints a separate report of one or more pages for each account number. Input to the printout program is the sorted "Hit" tape, and an account number address tape. Account numbers with no hits during search runs are identified, and the statement "NO HITS FOR THIS ACCOUNT AND PROFILE" is printed. If for some reason there are hits without matching account numbers, the report prints these out without an address, under the message "NO MATCH IN ACCOUNT FILE FOR THE FOLLOWING HITS."

5. Card-to-tape program and control card deck puts the account number addresses on tape. This can be done with any card-to-tape program.

Users have a wide latitude for expressing their requirements. However, search results will depend directly upon the information specialist's knowledge of the file structure, his skill in designing profiles, and his familiarity with appropriate terminology. For substructure searches, knowledge of WLN and chemistry is essential.

Neither software nor notation skill can replace imagination and ingenuity on the part of the user in designing intelligent questions.

The following examples illustrate how the WLN can be used for substructure searching by suitably arranged search questions.

A researcher interested in finding phenothiazines with piperazine in the side chain (Figure 8) will search for the string T C666 BN ISJ followed by the string T6M DNTJ or T6N DNTJ or T6K DNTJ.





A researcher interested in ethylamino steroids (Figure 9) will look for the string L E5 B666 followed by the stem 2N or 2M.

Successful answers demand two prerequisites: the storage of as comprehensive a file of chemical compounds as possible, and programs for interrogating and retrieving from the file.

The former is accomplished in *ICRS* by encoding new chemical compounds at an annual rate of over 160,000. The latter requires a continuing development of software. The foregoing descriptions indicate only a beginning along these lines. Work at ISI and elsewhere to develop programs for searching WLN files for a wide variety of uses is a continuing activity. Techniques used for selective dissemination will be quite different from those used for retrospective searches.

The *ICRS* RADIICAL system is designed primarily to provide substructure and other searches for monthly SDI reports. Depending upon requirements, each user may wish to make suitable modifications to take advantage of his own particular hardware configuration.

As part of the development of the *ICRS* system, ISI not only plans to issue annual cumulations of permuted notations, but also expects to enrich the tapes to include fragmentation codes used by many of the charter subscribers. Thus, it will then be possible not only to specify

Logic	Term	Type	Structure
1)	T C666 BN ISJ	STRING	
2)	FOLLOWED BY T6M DNTJ	STRING	
3)	OR T6N DNTJ	STRING	
4)	OR T6K DNTJ	STRING	

Query: Phenothiazines with piperazine in the side chain; that is, compounds encoded with the STRING T C666 BN ISJ, followed by the STRING T6M DNTJ, or T6N DNTJ, or T6K DNTJ.

Figure 8. Profile for WLN search on phenothiazines with piperazine in the side chain

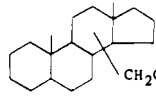
Structure	WLN	
	L E5 B666 . . . . . 2N 2M	
Logic	Term	Type
	L E5 B666	STRING
FOLLOWED BY	2N	STEM
OR	2M	STEM

Figure 9. Profile for ethylamino steroids

any combination of fragments, but also print out a line notation or a structural diagram.

ISI is now committed not only to develop a basic software package, but also to provide the necessary backup to implement its use.

#### LITERATURE CITED

- Revez, G. S., and A. Warner, "Retrieving Chemical Information with *Index Chemicus*," *J. CHEM. DOC.* 9, 106-9 (1969).
- a) Wiswesser, W. J., "A Line-Formula Chemical Notation," Thomas J. Crowell Co., New York, 1954.  
b) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw Hill, New York, 1968.
- "RADIICAL Users Manual," Institute for Scientific Information, Philadelphia, Pa., February 1969.
- Hyde, E., F. W. Matthews, L. H. Thomson, and W. J. Wiswesser, "Conversion of Wiswesser Notations to a Connectivity Matrix for Organic Compounds," *J. CHEM. DOC.* 7, 200-4 (1967).
- Fraction, G. F., J. C. Walker, and S. J. Tauber, "Connection Tables from Wiswesser Chemical Structure Notations—A Partial Algorithm," *Natl. Bur. Std. Tech. Note* 432, Issue September 1968.
- Kulpinski, S., N. London, D. Lefkowitz, and A. Genarro, "A Study and Implementation of Mechanical Translation from Wiswesser Line Notations to Connection Table," Volume I, October 25, 1967, 114 pp., and Volume II, November 30, 1967, 44 pp., Annual Reports to National Science Foundation on Contract NSF C-467.
- Finlay, A. C., private communication.
- Granito, C. E., J. E. Schultz, G. W. Gibson, A. Gelberg, R. J. Williams, and E. A. Metcalf, "Rapid Structure Searches via Permuted Chemical Line-Notations (III): A Computer-Produced Index," *J. CHEM. DOC.* 5, 229-32 (1965).