(16) Everitt, B. S. "Unresolved Problems in Cluster Analysis". *Biometrics* **1979**, *35*, 169–181.
(17) Williams, W. T.; Lance, G. N. "Hierarchical Classificatory Methods". In "Statistical methods for Digital Computers"; Enslein, K., et al., Eds.; Wiley: New York, 1977; Vol. III.
(18) Kowalski, B. R.; Bender, C. F. "Pattern Recognition. II. Linear and Nonlinear Methods for Displaying Chemical Data". *J. Am. Chem. Soc.* **1973**, *95*, 686–693.
(19) Everitt, B. S.; "Graphical Techniques for Multivariate Data"; Heinemann: 1978.

PULSAR: A Personalized Microcomputer-Based System For Keyword Search and Retrieval Of Literature Information

SCOTT F. SMITH,[1] WILLIAM L. JORGENSEN,*[2] and PHILIP L. FUCHS*[3]

Department of Chemistry, Purdue University, West Lafayette, Indiana 47907

Received March 9, 1981

A keyword-based storage and retrieval system for literature references has been developed with a TRS-80-II microcomputer. The system, called PULSAR, has been designed to provide and maintain rapid access to a personalized data base. Application of the PULSAR system to the literature of synthetic organic chemistry is described.

INTRODUCTION

When a scientist requires information, he will often draw first upon personal resources. These resources include all the specific contributions of that individual as well as any relevant literature information which he can recall. It is at this latter stage that major retrieval problems occur. A researcher acquires a rapidly increasing accumulation of data as his career proceeds. The specific manner in which these data are stored will determine their subsequent availability.

A variety of techniques have been traditionally employed for this task; most usually some variant of the well-known "card-file" system. In this case information is recorded on an index card, the file grows, and the cards are resorted as new categories are created. The system begins to become inefficient when the cards number in the thousands. At this point the categories have usually become too general, and it is apparent that substantial cross-indexing is necessary. The researcher might go through a temporary phase in which the cards can supposedly be sorted by passing a knitting needle through the edge of the cards. There are numerous inconveniences associated with all card-filing systems.

A number of general and extensive chemical information systems are now available to facilitate literature searching, including the Chemical Abstracts, Lockheed, NIH/EPA, and other systems.[4-7] Even with these systems readily accessible, there is strong justification for maintaining a personalized system restricted to the interests of an individual and based on the individual's own choice of keywords.[8,9] Due to the availability of microcomputers with diskettes, we have been able to develop such a system at reasonable cost capable of handling up to 20 000 references including keywords. The program is called "PULSAR" for Purdue University Literature Search and Retrieval system.[10]

PROGRAM FEATURES

The PULSAR system is implemented on a Radio Shack TRS-80-II computer with 64K bytes of memory, a printer, and from one to four 500K byte disk drives for a total storage capability of 2 Mbytes. The program is written in the BASIC language, and all 64K bytes of memory are used. The various options available in the PULSAR system are as follows:

I. Add Articles
II. Search for Keywords
III. Display Routines
(A) Display a single article
(B) Display a series of articles
(C) Display keywords alphabetically
(D) Display all journal book names
(E) Display system status information
(F) Display free space map
(G) Display disk directory
(H) Write a message on the printer
IV. Data Checking Routines
(A) Check keywords (start check at Keyword K)
(B) Check articles (starting with article No. *M*)
(C) Reformat (compact) link records
V. Editing Routines
(A) Edit keywords (rename/merge/delete)
(B) Combine keywords A and B to C
(C) Edit articles
(D) Edit journal names
VI. Disk File Management Routines
(A) Make a complete backup
(B) Format a new diskette
(C) Swap diskettes
(D) Move files from one disk to another
(E) Enable remote terminal

USING PULSAR

In the Add Articles routine (I), for each article reference, the following information is stored:

Entry number
Journal name
Volume, year, page
Type of article (Paper, Communication, Note, Thesis, Miscellaneous)
Keywords (1–8 keywords, ≤30 characters in length each)

The above information can later be retrieved by use of the Search for Keywords routine (II). For example, *A and B* will retrieve all article references containing both keyword A and keyword B. The logical expressions *or*, *xor*, and *not* are also usable; so *P and not* (*Q xor R*) *or* (*S and not T*) is a valid expression as well. Up to eight separate keywords may be referred to in any one search expression. Upon completing a search, the computer displays the number of matches upon a video screen. One may now elect to enter an alternate search

0095-2338/81/1621-0209$01.25/0

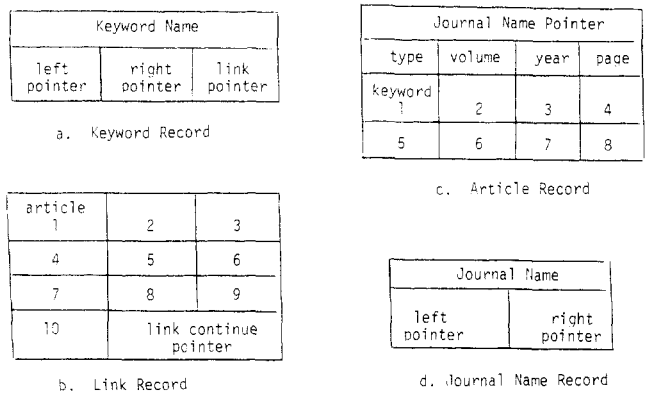

Figure 1. File record structures.

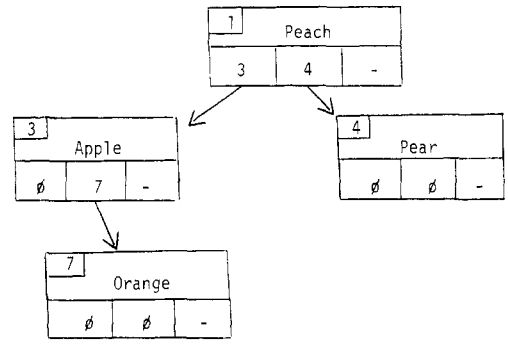

Figure 2. Sample keyword search tree.

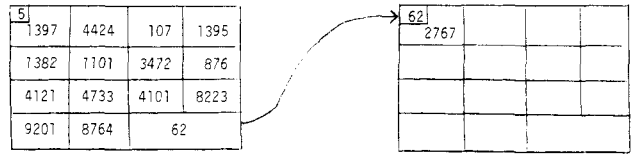

Figure 3. Link record continuation.

expression, to order all matches to be printed, or to display them one at a time on the video display. After viewing the display, one may subsequently choose to have this information printed out on a printer. Regardless of the specific search expression, *all information* which had been entered will be displayed. (This has the effect of giving a specific answer to a general question.)

The matches found in a given search may also be subsequently further qualified to be restricted to be from a specific journal, of a given type, within a range of years, or any combination of the above. This allows one to trim away extraneous information. Also, for a more interactive search, one may add or subtract from the matches found in the previous search. To do this, the character "@" represents that which was found on the previous search, so the expression @ and U would match all of the articles found in the previous search that additionally have the keyword U.

Various information may also be displayed and/or printed, including a given entry (III-A), a series of entries (III-B), an alphabetically arranged keywork list (III-C), the journal list (III-D), system status information (III-E), the free space map for any drive (III-F), the disk directory for any drive (III-G), messages written to the printer (III-H).

There is a self-correcting data check program (IV-A,B) which also allows easy maintenance of the system as well as the ability to monitor the number of articles which reference a given keyword.

Editing routines that allow renaming or deletion of keywords within the keyword list (V-A) are available. There is an option which allows the combination of two keywords to create a new "combined keyword" (V-B). It is also possible to edit any or all information within each individual entry (V-C) as well as to correct the journal names (V-D).

There is a set of disk file management routines which enables periodic copying of the data base to ensure the security of the data (VI-A). There are also additional options for system maintenance and modification (VI-B-D).

FILE STRUCTURE

The program uses five files to store the data. Each of the files is divided into records of equal length, and the records may be read and written in any order, i.e., they are random-access files.

The keywords are stored in a separate file, with a record structure as shown in Figure 1a. Each record contains a keyword name of up to 30 characters in length. The keywords are stored in a binary search tree format for efficient searches, additions, and deletions. One record is designated the "head" of the tree. The left pointer of each keyword contains either the record number of a keyword alphabetically preceding it or a "0"; the right pointer is for a keyword following it alphabetically. The result of such a formulation is an easy method for searching for keywords; the structure also facilitates additions and deletions of keywords. For example, in Figure 2, in order to search for "Orange", the program first checks the head of the tree. "Peach", the keyword with record number 1 (the number in the upper left-hand corner indicates the record number), is found. Since Orange alphabetically precedes Peach, the record number contained in the left pointer of Peach is checked (record 3). Orange follows "Apple"; therefore, the program will then check the record number in the right pointer of Apple (record 7) which is Orange, and thus the keyword is located.

In order to add a keyword, the program searches alphabetically until a 0 is found in either the left or right pointer. That zero is changed to the record number of the keywords being added. Deletions are more complex; however, they can still be performed in a minimal amount of time.

The links pointer contains the number of a record in the link file where the article numbers associated with the keyword are stored.

The link file contains records which each hold 10 different article numbers (Figure 1b). For more than 10 article numbers referring to the same keyword, additional link records can be linked together via the link continue pointer (Figure 3). This allows the system to be as dynamic as possible, i.e., keywords can be added and deleted easily, and the number of references for each keyword can grow gradually without a known upper limit ever being specified. Internal limitations, however, currently require a maximum of 1000 references allowable for each keyword.

The article file is used to store information pertaining to each article; each record of the file contains all information about a single article (Figure 1c). This includes pointers containing the record numbers in the keyword file of the keywords which refer to the article. This is primarily for display purposes—the keywords used in a given article are always displayed when the article is displayed. The volume, year, page, type, and a pointer containing the number of a record in the journal file are stored as well. Rather than storing a 40-character journal name with each article, a number referring to a record number in the journal name file is used. In this way each journal name is stored only once; thus disk space is conserved.

The journal name file contains the journal names, which are stored in a binary search tree format similar to the format of the keywords, but without the link pointer (Figure 1d).

The free record file stores record numbers which are made available when a keyword is deleted. When a new keyword

```
# 1 - JOURNAL NAME - CITRUS WORLD
TYPE - PAPER
VOLUME 3   YEAR 74   PAGE 95
KEYWORDS:
1. ORANGE            2. TANGERINE
3. GRAPEFRUIT        4. CHEMICAL CONTROL
5. INSECT CARRIED DISEASE
```

```
# 2 - JOURNAL NAME - PRODUCE DEALERS DIGEST
TYPE - COMMUNICATION
VOLUME 104   YEAR 81   PAGE 1027
KEYWORDS:
1. CITRUS FRUIT      2. ORANGE
3. GRAPEFRUIT        4. PACKAGING
5. INTERSTATE COMM.
```

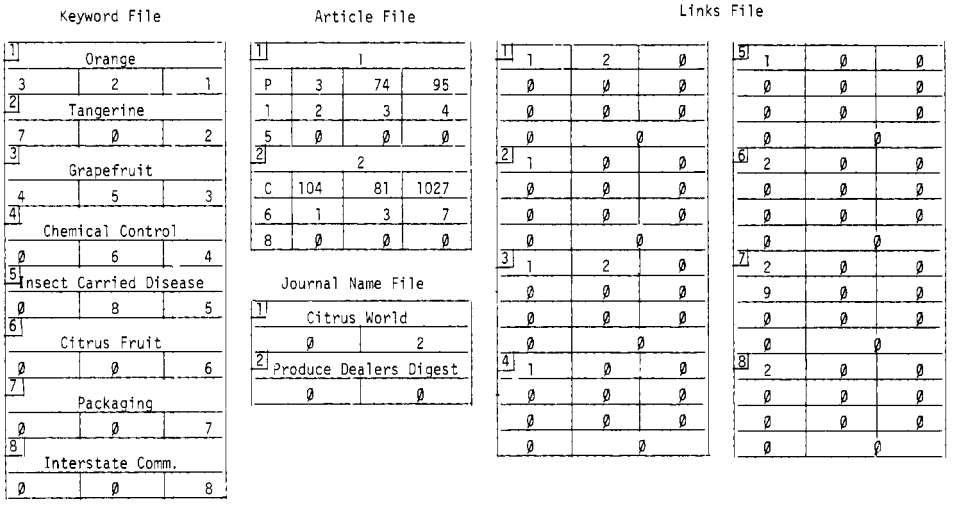

Figure 4. (a) Two sample articles. (b) Example of the file structure after the addition of two articles.

Table I

routine name	parameters in	parameters out	comments	options where used
keyword search routine	keyword name	record number in keyword file	the record number of the given keyword is found by searching through the keyword tree; if the keyword is not found, a flag is set	II; IV A, B; V A, B, C
reference fetch routine	keyword name	a list of the articles containing the keyword	the article numbers are retrieved from the links file and stored in the computer memory for fast accessing	II; IV A, B; V A, B
keyword add routine	keyword name article number	record number where the keyword was added	the keyword is added to the tree structure	I; V A, B, C
reference add routine	keyword name article number		the article number is added to the list of references for the keyword (i.e., the number is put in the links file list for the keyword)	I; IV B; V B
reference delete routine	keyword name article number		opposite from above, the article number is removed from the list for the keyword	V C
keyword delete routine	keyword name		the keyword is deleted from the tree, and its record number is placed in the free records file	V A, B, C
journal search/add routine	journal name	record number of the journal in the journal name file	the record number of a given journal name is found by searching through the journal tree; if the journal does not exist in the tree, it is added	I; V D

is added, it will be stored in one of these free spaces. If there are no free records available, as will most often be the case, the keyword is added to the end of the file.

Figure 4 contains an example of how the file structure would appear after two articles have been added. Figure 4a contains the information as it would be entered in the add articles option (I), and Figure 4b illustrates how the information would be represented in the file structure.

PULSAR contains a number of subroutines which perform the file manipulations. A brief description of each routine appears in Table I.

APPLICATIONS OF PULSAR TO SYNTHETIC ORGANIC CHEMISTRY

The synthetic organic chemist faces a special problem in dealing with the scientific literature. The very nature of the discipline demands exceptionally broad coverage of the literature. During the course of a total synthesis for example, the practitioner may require information about virtually any facet of organic chemistry.

The PULSAR system offers considerable assistance in the information handling ability of the practicing scientist. There are two compelling reasons for adopting such a system. (1) Since it is a *personalized* system it is perfectly matched to the thought processes of the individual user. As a keyword system it is by its very nature programmed to respond in the *exact language* of the individual user. It is this specific ability that makes it more responsive than systems which are more general in nature (Chemical Abstracts, Lockheed, etc.). The purpose of the PULSAR system is not to replace those systems which presently provide extensive access to the literature but rather to provide an organizational framework upon which to develop

an individual's own perspective of the literature. (2) In the process of deciding which keywords to assign to any given article, the reader of the literature is forced to *carefully specify* how each particular article fits into the current scientific dicipline. This individual responsibility results in more careful reading of the literature and is a more subtle reason for the use of PULSAR.

PULSAR IN OPERATION

When an article is read, up to eight keywords are chosen which outline its scientific content. In those rare instances (~1% of the 5700 articles currently entered) where more than eight keywords are necessary to fully describe a given article, a second set of keywords is selected, and the article is referred to a second time.[11] Keyword order is designated so that a crude narrative can be inferred from examination of the keyword sequence. The importance of this practice can be seen in example 1. This article refers to the 1,4 addition of a metalated benzyl nitrile to a vinyl ester and not to the metalation of a vinyl ester.

Example 1

#2227—JOURNAL NAME—TET LETT
TYPE—COMMUNICATION
VOLUME 79 YEAR 79 PAGE 1201

KEYWORDS:

1. BENZYL NITRILE	2. METALATION
3. ALPHA-NITRILE ANION	4. 1,4 ADDITION
5. VINYL ESTER	6. TOTAL SYNTHESIS
7. ! DAUNOMYCINONE	8. * KENDE A

Example 1 also demonstrates the use of "punctuation" to subalphabetize the keyword list. All natural products are specified by an exclamation point prefix and all authors by a starred prefix. This practice allows the "natural product" or "author" keywords to be displayed in their entirety (III-C).

The most critial determinant in optimizing the system efficiency lies with careful selection and maintenance of the keyword data base. In its current configuration (PULSAR 4/3)[12] the system will store ca. 20 000 articles and will have a base of approximately 3000 keywords. In the initial phase of developing the data base, the keyword file grew rapidly. After approximately 3000 articles had been added, the keyword file began to expand far more slowly and is now easily maintained between 2500–3000 keywords by several of the interactive features of the PULSAR system.

There are five features which allow the keyword file to be maintained at a relatively constant size: (1) The ability to print or display the entire keyword list with or without the number of articles referenced (III-C) enables scanning of the entire list in order to find errors and identify redundancies (i.e., BETA-HALO ALCOHOL ≡ HALOHYDRIN ≡ BETA-HYDROXY HALIDE) for further editing. (2) An automatic printout of any new keyword which has been created during the process of entering articles/keywords to the data base enables the user to reconsider the creation of each new keyword as well as to screen for misspellings. (3) An editing program (V-A) allows keywords to be changed, merged, or deleted from the keyword list. (4) The alphabetical keyword list routine (III-C) displays keywords according to the number of articles in which they appear. For example, all the keywords which have a number of articles fewer than N or more than M may be displayed. This allows for the identification of underused or misspelled keywords. This also determines which keywords are overused (the system has an upper limit of 1000 articles/keyword). (5) In the case of keywords which are overused and are therefore probably too general, there is an editing program which allows the creation of combined keywords (V-B).

For example, a keyword check shows VINYL ESTER to be listed in 145 articles (of 5700 entered) while 1,4 ADDITION occurs 255 times. A search of 1,4 ADDITION and VINYL ESTER (example 2) shows 44 matches.

Example 2

ARTICLES 5700, MATCHES FOUND = 44
SEARCH EXPRESSION = 1,4 ADDITION AND VINYL ESTER

In this instance it would be possible to use the editing routine (V-B) to create the new keyword 1,4 ADDITION VINYL ESTER. In this routine the user may elect to retain or delete the second keyword (the first keyword, 1,4-ADDITION in this example will be replaced by the merged keyword whether or not the second keyword is retained). It is important to recognize that not every mutual occurrance of 1,4 ADDITION and VINYL ESTER would qualify as a candidate for the merged keyword 1,4 ADDITION VINYL ESTER. Entry 2227 of Example 1 could be reasonably changed to example 3.

Example 3

#2227—JOURNAL NAME—TET LETT
TYPE—COMMUNICATION
VOLUME 79 YEAR 79 PAGE 1201

KEYWORDS:

1. BENZYL NITRILE	2. METALATION
3. BENZYL ANION	4. 1,4 ADDITION VINYL ESTER
5. TOTAL SYNTHESIS	6. ! DAUNOMYCINONE
7. * KENDE A	

However, articles 2147 and 438 (example 4) indicate that the vinyl ester is a *product* of a 1,4-addition reaction rather than a substrate, and one would not want to change them. To accommodate this situation the combine keywords editing routine (V-B) asks the operator to specify whether or not each individual candidate will be merged.

Example 4

#2147—JOURNAL NAME—SYN COMM
TYPE—NOTE
VOLUME 9 YEAR 79 PAGE 325

KEYWORDS:

1. 4 RING	2. BETA'-X VINYL ESTER
3. X=ALKOXY	4. CUPRATE
5. 1,4 ADDITION	6. SN2'
7. ADDITION-ELIMINATION	8. VINYL ESTER

#438—JOURNAL NAME—CR ACAD SC PARIS C
TYPE—COMMUNICATION
VOLUME 267 YEAR 68 PAGE 738

KEYWORDS:

1. BETA-X VINYL ESTER	2. X=HALO
3. CUPRATE	4. 1,2 VS 1,4 ADDITION
5. 1,4 ADDITION	6. ADDITION-ELIMINATION
7. VINYL ESTER	* NORMANT H

The principal function for which the PULSAR was created was the efficient recovery of literature information. An example of a keyword search is shown in example 5.[13]

Example 5

ARTICLES 5700, MATCHES FOUND = 9
SEARCH EXPRESSION = (CUPRATE OR ORGANOCOPPER OR GRIGNARD OR ORGANOLITHIUM) AND ADDITION-ELIMINATION AND (VLGS ACID HALIDE OR

VLGS ESTER OR VLGS THIOLESTER)
#2950—JOURNAL NAME—J ORG CHEM
TYPE—COMMUNICATION
VOLUME 44 YEAR 79 PAGE 3437

KEYWORDS:
1. TIN ANION
2. MIXED
3. CUPRATE
4. ADDITION-ELIMINATION
5. VLGS ACID HALIDE
6. BETA-X ENONE
7. X=STANNYL
8. * PIERS E

#1732—JOURNAL NAME—J ORG CHEM
TYPE—COMMUNICATION
VOLUME 40 YEAR 75 PAGE 2694

KEYWORDS:
1. 6 RING
2. VLGS ACID HALIDE
3. MIXED
4. CUPRATE
5. ADDITION-ELIMINATION
6. ENONE
7. * PIERS E

This particular search found nine matches (two of which were selected to be displayed by the operator). The entries are always displayed in the order of decreasing entry number. Since this is a dynamic system where the keywords for current literature articles are constantly being added, this feature will generally display the most current information first.

A specific problem in using a keyword-based system for literature retrieval in organic chemistry is associated with the variability of organic nomenclature. There is much personal latitude in the use of hyphens and spaces in designating organic functional groups. For example, BETA-KETO-ESTER, BETA-KETOESTER, BETA KETO ESTER, and BETA KETOESTER may all be used to represent a single functional group. Since the computer would normally perceive each of these as a distinct keyword, an additional routine was developed which removes all hyphens and spaces from keywords during addition, search, and editing procedures. However, the keyword is retained and displayed in the form first entered into the data base. Although the "compacted" form is never seen, the net effect is that the program prevents accidental redundancies from occurring.

In conclusion, it is felt that the PULSAR system provides a highly versatile tool for the organization and retrieval of literature information and should be also applicable in many areas outside of organic chemistry.

REFERENCES AND NOTES

(1) Purdue University Undergraduate Research Associate.
(2) A. P. Sloan Fellow, 1979–1981; Camille and Henry Dreyfus Teacher-Scholar, 1978–1983.
(3) A. P. Sloan Fellow, 1977–1979.
(4) Milne, G. W. A., *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 204.
(5) Roush, P. F.; Seitz, J. T.; Young, L. F. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 73.
(6) Ziegler, H. J. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 141.
(7) McCarn, D. B.; Leiter, J. *Science (Washington, DC)* **1973**, *181*, 318.
(8) Van Ree, T., *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 152.
(9) Wilde, D. V.; Starke, A. C. *J. Chem. Doc.* **1974**, *14*, 41.
(10) Copyright 1981, Purdue Research Foundation. The PULSAR program is commercially available from Litindex, Inc., P. O. Box 2274, West Lafayette, IN 47906.
(11) The practice of one of the authors is to simply write keywords on the specific article as the current journal is being read. The article is then copied, a number stamped on it with a consecutive numbering machine, and it is stored in its entirety in a 3-ring binder. Entry of new articles to the data base can be performed by anyone familiar with the operation of a typewriter keyboard.
(12) This designates a four-drive system where three drives are used for storage and one drive is available for copy purposes.
(13) Vlgs is an abbreviation for vinylogous.

Present Status of Inorganic Chemical Nomenclature†

W. CONARD FERNELIUS
Kent State University, Kent, Ohio 44242

Received June 26, 1981

A systematic chemical name is one which portrays the essential structural features of a chemical compound by some general pattern. For most purposes it is unnecessary to write nomenclature rules in such detail as to provide a single name for each compound. Like all human activities nomenclature patterns change with time. This is essential to meet new conditions, to secure greater generality, or to obtain simpler names. While nomenclature specialists must be attuned to the needs of nomenclature users, their suggested solutions to be successful must be acceptable to the users. The presentation details with particular reference to inorganic chemistry (1) the committee–commission structure in this country and internationally, (2) significant accomplishments in the past half-century, (3) developments in progress, and (4) areas where nomenclature developments are needed.

Although the nomenclature of chemical compounds may appear confusing, if not meaningless, to the uninitiated, it is amazingly exact and simple in concept. Chemistry has benefited throughout its history by the nomenclature principles adopted by the early pioneers of the science.[1] They realized the importance of systematically relating names to the composition of the individual compounds. Their system with relatively minor modification served inorganic chemists well into the present century. Organic chemists early encountered marked inadequacies in the original systematic nomenclature and shifted to structure, rather than simple stoichiometry, as the basis for nomenclature. The adoption of structure as the basic consideration for names of inorganic compounds followed later but is now firmly established.

Many have asked, "Since structure is the important criterion for naming chemical compounds, why not dispense with names

† Adapted from an address on the occasion of the presentation of the Patterson–Crane Award of the Dayton and Columbus Sections of the American Chemical Society, May 22, 1981, Dayton, OH.

0095-2338/81/1621-0213$01.25/0