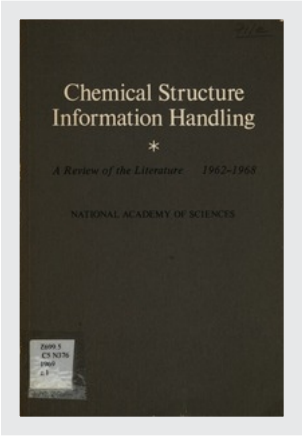


Chemical Structure Information Handling: A Review of the Literature, 1962-1968

DETAILS

144 pages | 5 x 9 | PAPERBACK
ISBN 978-0-309-36506-2 | DOI 10.17226/21566

AUTHORS

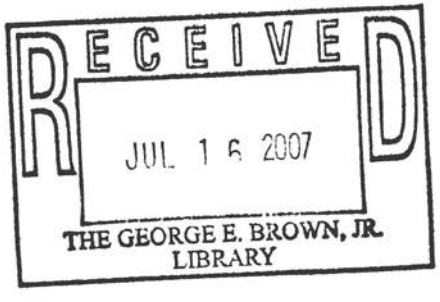
Bart E. Holm, Chairman; Committee on Chemical Information; Division of Chemistry and Chemical Technology; National Research Council

Chemical Structure Information Handling

*

A Review of the Literature 1962–1968

COMMITTEE ON CHEMICAL INFORMATION
DIVISION OF CHEMISTRY AND CHEMICAL TECHNOLOGY
NATIONAL RESEARCH COUNCIL



NATIONAL ACADEMY OF SCIENCES
WASHINGTON, D.C.
1969

Project supported in part by the Army Research Office
under Contract No. DA 49-092-ARO-58
and in part by a grant from the American Chemical Society.

Publication 1733

Available from
Printing and Publishing Office
National Academy of Sciences
2101 Constitution Avenue
Washington, D.C. 20418

Library of Congress Catalog Card Number 70-602073

PREFACE

*

The Committee on Chemical Information is part of the Division of Chemistry and Chemical Technology of the National Research Council (NRC). The committee's broad function is to survey and report on significant technical trends in needs, sources, usage, and methods in the utilization of chemical information. A specific function of the committee is to act as a clearinghouse for assembling, analyzing, and disseminating knowledge about significant new developments.

In 1964, the committee issued Survey of Chemical Notation Systems*, which covers significant practices in the use of nonconventional chemical notations in the United States through 1961. In 1965, the committee issued a second report, Survey of European Non-Conventional Chemical Notation Systems†, which deals with significant European practices in 1962. Since 1962, important additions to the technology and new applications have been described in a multitude of publications. Therefore, a new survey of chemical structure information handling was needed to bring the earlier reports up to date.

The earlier reports were based on actual interviews with organizations doing significant work in chemical structure information handling. The present review is based on a different approach, utilizing the published and unpublished literature as input, augmented by the personal communications and experiences of the committee's

*National Academy of Sciences Pub. No. 1150.
†National Academy of Sciences Pub. No. 1278.

membership. Furthermore, the scope has been broadened somewhat to include additional aspects of chemical information handling, as outlined in the Introduction.

Necessarily, boundaries were drawn setting limits to the objectives of the review. In general, candidate material included developments in the handling of chemical structure information, such as translation of linear notations to connection tables. Excluded from consideration were discussions of general, widely applicable information methods, even when such methods were reported as applicable to the handling of chemical information. On this basis the report does not include such topics as current-awareness and selective-dissemination applications, computer-based editing and publication, error prevention in the literature, microforms, and behavioral patterns of users. Two particular topics set aside as beyond the objectives of the review were developments in traditional chemical nomenclature and sources of chemical information.

The objective of this review was to put into focus the various developments that have been reported in the field during the past few years. Effort was devoted to bringing together programs and activities of similar intent and technical content, even though these activities may have been separated by time and distance and may have been reported in different sources. Attempts were made also to summarize the highlights of large-scale projects of long range that have been reported extensively but the goals of which may have been obscured by interim documentation.

Many authors give estimates of operational costs. These cost estimates are usually not comparable with others because of differences both in the quality of drawn structures and in the classes of compounds covered. They are dramatically affected by accounting methods, which seldom are stated. The committee considered omitting cost data altogether because of the difficulties of meaningful comparison, but decided to include them when reported. They should be viewed with extreme caution, however.

Some special contributions deserve to be acknowledged. Partial support for this report and other work of the committee has been received by the National Academy of Sciences from the Army Research Office and from the American Chemical Society. The project was conceived, designed, and initiated by James M. Mullen of the Shell Development Company, whose untimely death is deeply regretted. The report was drafted by Madeline M. Henderson, consultant to the committee, with assistance from Beatrice Marron and Mary Jane Ruhl. Completion of the project was directed on behalf of the committee by Herbert R. Koller of Leasco Systems and Research Corporation. All members of the committee participated actively in the

preparation of the final draft, contributing input, comments, advice, and criticism in creating what we hope is a useful state-of-the-art report of chemical structure information handling covering the period from 1962 until late 1968.

Bart E. Holm
Chairman
Committee on Chemical Information

CONTENTS

*

I

INTRODUCTION

*

A. CHARACTERISTICS OF THE CHEMICAL LITERATURE

Because of the nature of chemistry and its records, chemical research-and-development activities demand frequent access to the literature of the field. But that literature is beset by the same problems vexing the literature of all scientific and technical fields, namely increasing volume, proliferating areas of specialization, and the need for cross correlation with other disciplines. Although the handling of chemical information and its organization for later search, in general, involve the same basic information-processing techniques that are used in other disciplines, this field is unique in that its literature has as a least common denominator chemical structure information, which can be used as a focal point for storage and later search.

In attempting to cope with problems of chemical documentation, chemists most frequently feel the need for reference to chemical formulas and structures as methods of entry to the literature of the field. "Although there are many other aspects of chemistry, the chemist is usually concerned with the molecular identity of materials, and it is this kind of subject matter about which he most frequently communicates."[101] Synthetic processes, purification procedures, biological effects of pharmaceuticals, physical and chemical properties—these are meaningful only when associated with particular compounds or classes of compounds and their structures. The means for designating or labeling chemical compounds and identifying

their structures are therefore extremely important. Chemists desire detailed understanding of the structural characteristics of chemical compounds and materials in order to search out specific compounds, groups of related compounds, or particular activities associated with structural features.

The discipline of chemistry enjoys highly organized services and tools to accomplish its needed access to the literature. For example, the worldwide fame of Chemical Abstracts Service (CAS) as a source of notification of chemical progress is well founded. In a British survey of information needs of physicists and chemists,[1] the chemists evidenced more regard for abstracts than did the physicists. ". . . the central position held by Chemical Abstracts . . . has no counterpart in physics. . . ." Abstracts and original papers are considered the most important sources of specific information, while meetings and conferences are the best vehicles for "current awareness." Reviews also are widely used, both for keeping up-to-date in one's specialty and for learning about new fields. In the survey just mentioned, a strong desire was expressed for more reviews, and particularly for more introductory reviews. The same opinion was expressed in a Symposium on Critical Reviews at the April 1968 national meeting of the American Chemical Society (ACS).[146]

In handling the chemical literature, the description of chemical compounds is of paramount interest, and the structural diagram has become the basic medium of communication. "The structural diagram is a two-dimensional projection of molecular structure."[40] It is also "the simplest and most familiar means for representation and communication between chemists."[144]

To complement the structural diagram, systematic nomenclature has been developed as an essential part of the communication system. As trivial names proliferated, it became clear that systematically derived names were needed to improve communication. Rules provide for systematic numbering of atoms in the structural formula, followed by generation of a name with standard prefixes, stems, and suffixes, combined with numbers and, occasionally, Greek letters. After sufficient exposure to such rules, the average chemist can usually develop a reasonably descriptive name that will serve adequately for discussing a given structure; but he has no confidence, especially if the compound is a strange or moderately complex one, that an indexer (or for that matter, another chemist) would use the same name. That is, "presently used systematic nomenclature is based directly on the diagrammatic language and is therefore no more exact than the diagrammatic language itself."[181] The verbal counterparts of structural diagrams can provide easier recognition of equivalency than the diagrams themselves. However, for some

classes of compounds, especially new ones or those only recently attracting interest, nomenclature rules may be slow to develop and may be obscure. In such cases, more-or-less uncontrolled nomenclature is introduced into the literature.

Naming of compounds is necessary for oral communication among chemists, but the inadequacy of names as an indexing tool is becoming increasingly evident as the quantity of chemical information grows. Both structural diagrams and systematic nomenclature aid in identifying specific compounds, and names can be suitable for indexing when they provide specific addresses for compounds so that they can be located again. But names provide a poor basis for the classification or grouping of compounds that have similarities in structure. "The information required to answer a question may be accessible through the index, but it often involves too much human effort to seek out those kinds of detail which are not high in the established order of the classical index's hierarchical priorities."[215] Difficulties with traditional methods, together with the need for more flexible methods for handing large numbers of compounds, have stimulated the development of alternative techniques for representing structures. Significantly major developments and uses for structure coding have evolved in pharmaceutical and medical research organizations, where comparative evaluation of large numbers of complex compounds is a constant requirement.

B. IDENTIFICATION OF CHEMICAL STRUCTURES

Techniques for identifying structural features by numeric or alphabetic codes were developed along two major lines: fragmentation techniques and notation techniques. In a fragmentation code," . . . pre-selected structural aspects of compounds [are represented] by identifying atoms or groups of atoms (fragments) that have significance for later searching."[59] These fragments are recorded on equal terms, but their context is lost. The original structure usually cannot be regenerated. However, such codes have proved to be adequate and practical for many purposes.

A notation is a "unique code for the structure of a compound which uses a group of letters and numbers that can be written essentially on one line."[59] Notations are designed to represent complete two-dimensional descriptions of three-dimensional structures. Notation systems here refer to systems designed to be used by chemists, with manual or automated procedures. Thus, notation systems attempt to "retain the majority of the element symbols already familiar to the chemist."[70] If new symbols are introduced for fre-

quently encountered atoms and functional groups, they are often mnemonically suggestive. Such notations can be organized and printed in list form to make a useful index, which then can be searched manually or by computer. Notations can also be permuted (the order of their symbols interchanged) to make a multiple-entry index, just as titles are permuted to make a keyword-in-context index.

Work aimed toward establishing a major system for handling chemical information requires the ability to designate chemical structures precisely and without recourse to structural formulas. An adequate information-handling system must accomodate not only all known structural types, but also structural types on which there has merely been speculation. To this end Tauber proposed an augmentation of the Cahn-Ingold-Prelog (C-I-P) rules for designating absolute configuration and the addition of a new (specialized) set of rules to cover a wider range of molecularly dissymmetric structural types.[183] The C-I-P rules claim to cover all known types of dissymmetry deriving from tetracovalent and tricovalent atoms.[32] The molecularly dissymmetric structures for which Tauber suggested further rules are of two types: (1) substituted catenanes, which are stereochemically related to substituted allenes and spiro compounds, and (2) knots, compounds with dissymmetry due entirely to topology, independent of substitution. The C-I-P convention is inapplicable to knots and, before it can be applied to catenanes, requires a further convention for defining the "near groups."[183]

The performance of a structure information system in a practical application depends upon several factors, including use (that is, the nature of the job to be done) and the principles underlying the design of the system. Two completely distinct and exclusive functions to be performed by the system have been identified as "description, definition, or delineation of a structure, and . . . organization of structures for indexing purposes."[19] These are also the functions commonly expected of systematic indexing by name.

While the notation form representation of the structure of a chemical compound may be thought of as a "name," use of the notation in lieu of the conventional structural formula, systematic name, or trivial name in all the ways these expressions have been used is not generally feasible. The use of a notation, for example to convey structures in speech, appears to be less practical than in writing. "Even in writing, chemists probably would find notations difficult to follow when substituted for structural formulas in reaction flow charts."[19] Since the functions of delineation and indexing are distinct, a technique for serving one function may not serve the other well. The structural formula, for example, does well in delineating a structure, but does not function as well as a basic indexing tool.

Consequently, the designer of a system must choose which of the functions he wishes to emphasize. The relative emphasis given to the delineation and to the indexing functions in the design of a structure information system may be influenced by the use and manner of use intended by the designer.

One of the problems faced in the introduction of a notation system is that it represents a new "language" that the user must learn. However, in notation systems, rules are kept to a minimum and are generally simpler than those for nomenclature. Because of these characteristics, a notation system can be learned quickly and used easily. Interestingly, it is usually "easier to learn how to decode a cipher, i.e., to derive the structural formula from the cipher, than it is to learn to cipher correctly."[101] This, of course, is also characteristic of systematic nomenclature, as suggested above. Nomenclature and the principles of notation schemes must be taught if chemistry students are to learn adequately how to handle their chemical information problems.

"Substructure information in notations is carried in the cipher symbols and their combination; generic retrieval is accomplished through symbol comparison in a manner analogous to searching via fragment descriptors."[87] The dependence on hierarchical rules and on an intricate symbol language analogous to the customary systematic nomenclature is a handicap; however, the "most serious limitation is that a cipher is a condensed statement of the structure. The condensation required prejudgment as to what will be of interest in the way of chemical groupings to future generations of chemists, and even to those at present whose field is widely removed from that of the originator of the cipher."[73] This deficiency is neutralized to some extent by the possibility of generating an atom-to-atom connection table from the cipher by way of a computer program and a topological code.

C. COMPUTERS IN CHEMICAL INFORMATION HANDLING

"It is natural that those working in the information field would begin to investigate the capability [of the computer]."[20] Just as the availability of punched-card equipment influenced the development of the fragmentation and notation techniques, so the increasing availability and acceptance of the computer as an information-handling machine has opened further possibilities. Topological codes are being used as structure descriptions showing atom-to-atom connections, with the atoms as nodes and the connections as branches of a network. In a sense, such a code "is a mathematical snapshot of the complete

structure, and the corresponding matrix representation is a numerical analog and a structural model which can be handled via computer programs."[87] The theories and techniques of network analysis that have evolved in recent years provide the basis for several systems using matrices to encode structures.

In the practical systems that have been developed, generic information is stored within the topological structure code, and retrieval is determined by the substructure selected for the search. In contrast to other methods, the degree of discrimination is optional and is determined by the choice at the time of search, not by a prior structure dissection at the time of indexing. Thus, in concept, the topological description approaches a maximum in coding technique: a unique, complete description of structure and a flexible, unlimited substructure classification. The code can provide a versatile method for both identification and generic classification.

The capabilities of computers for information processing have fostered developments in total systems as well as in techniques for identifying structures. The Chemical Information Program (CIP) of the National Science Foundation (NSF), supported on an interagency basis by the federal government, was established to "modernize existing chemical documemtation services and to develop new concepts and tools for handling chemical information on a large scale."[191] The basic contract with the American Chemical Society, in particular CAS, during the early years of the program (1965–1966) called for development of a large registry system for filing information about chemical compounds and creation of a file of sufficient size to test many problems of utility, serviceability, and economics of the operation of such a system on a national scale.[143] Much of the effort has gone into developing systems for handling chemical structure that are adaptable to mechanization.

Chemical Abstracts Service contracts with the NSF in 1967 included support from the Food and Drug Administration (FDA) and the National Library of Medicine (NLM). An activity of growing importance was the exploration of potential areas of cooperation between CAS and other processors and users of chemical information. Significant progress in this direction was made in 1967 in a joint effort by FDA and NLM to use the CAS Registry and substructure search techniques as a common basis for assignment of index terms and coordination of nomenclature and literature references for chemical substances.[7] The CAS/FDA/NLM cooperative arrangement should "produce valuable experience and highlight problems of general applicability."[154]

In addition, the Chemical Society (London) established an information center at the University of Nottingham to adapt CAS computer-

based services to the needs of British scientists and engineers and continued a program of reciprocal internships through which staff members of each organization are working in the other's information program. In 1967 "a representative of the Karolinska Institute of Sweden joined the CAS staff to gain experience in various phases of computer information-handling operations [and] discussions and reciprocal exchange visits with German chemical information groups also continued."[7]

It has become obvious that the objectives of the CIP cannot be divorced from the multifold computer activities and plans of federal agencies, of industrial firms, and of leaders in other scientific disciplines. The interrelationship is brought out in a survey of federal chemical information and data systems conducted for NSF by the John I. Thompson Co. to collect "basic data for use in the planning of a National Chemical Information System."[152] Information Management, Inc., made another study for NSF of plans and specifications for a national chemical information system.[124,125] Science Communication, Inc., studied chemical data-compilation activities at ten major centers and made an exhaustive analysis of many data services in chemistry to determine the degree of emphasis to be placed upon data compilations in the design of a national system.[28] Of course, examination of the needs of chemists must not be limited to those who practice chemistry or chemical technology alone. "Medical, biological, physical scientists and others, have information needs in chemistry. . . . By the same token, people whose primary interests are chemical have information needs in other disciplines."[215]

Many problems remain to be solved in the realm of chemical information handling. Among them are: (1) automatic analysis and coding of the textual and diagrammatic materials that are found in the chemical literature, (2) better knowledge of the chemists' specific needs for information in a wide variety of environments, (3) development of "conversational" systems that guide the searcher to improve his questioning technique through partial and intermediate answers produced by the machine system, and (4) improved techniques for comparing and evaluating the economics and utility of different information systems and techniques.[101]

These are "problems that exist as general problems in information retrieval, but each has special importance for chemical information systems. [In addition] four problems are peculiar to the chemical information field."[101] These are (1) computer composition of structural diagrams for transient display and for preparation of graphic-arts materials, (2) the handling of information dealing with inorganic chemistry (the small amount of work produced in this

area to date is, in general, less advanced than that produced in organic chemistry), (3) classification approaches commensurate with the logical characteristics and capabilities of computers, and (4) methods for adequately correlating structural features of materials with functions and properties and for predicting structures having specified properties and uses.

D. SCOPE OF THIS REPORT

Since the appearance of the National Academy of Sciences–National Research Council reports, Survey of Chemical Notation Systems (1964) and Survey of European Non-Conventional Chemical Notation Systems (1965), the literature dealing with chemical information handling has burgeoned impressively. Those two publications were based on first-hand study of on-going projects and operating installations. They were not concerned primarily with the literature of their subject areas, but major references up to 1962 were cited in them.

The present report reviews the published literature from the period of those two reports through the spring of 1968. Specifically, it includes the papers presented at the April 1–5, 1968, symposium on Chemical Notation Systems sponsored by the Division of Chemical Literature of the ACS. A very few more recent references are also covered.

The word "published" in the context in which it is used here includes technical reports as well as journal articles, books, and published proceedings. It also includes papers presented at professional society meetings that were distributed as preprints or as copies made available by the authors.

As the literature was analyzed, it fell into several categories of specific discussion. The flow and interrelationships of these categories are reflected in the Contents. The arrangement reflects the interests and viewpoints of the members of the Committee on Chemical Information. In addition to documents pertaining to particular categories, there are several of broad scope that serve as background or summaries of the entire subject of the report. For example, worldwide chemical information facilities were extensively surveyed by the NSF[145] as a contribution to the study jointly sponsored by the International Council of Scientific Unions (ICSU) and the United Nations Education, Scientific, and Cultural Organization (UNESCO) of the feasibility of a worldwide scientific information system. The report covers the full range of services from traditional publication through the developing computer-based systems. The survey is

"concerned with documentation rather than oral communication and with recognized systems . . . rather than informal systems. . . ." No attempt was made to evaluate, although illustrative comparisons were made; the overview is descriptive in nature. An extensive bibliography and a glossary of terminology were appended.

Lynch reviewed and compared various representations that have been suggested for describing the topology of chemical molecules[120]; his paper includes a good bibliography. Developments in chemical information handling are examined broadly in the chapter by Tate entitled "Handling Chemical Compounds in Information Systems"[181] appearing in the Annual Review of Information Science, Volume II (1967). Pertinent developments of 1966 are covered, as well as major developments prior to 1966. A most useful bibliography is included. Other documents of a general or review nature have been cited in the present review in the sections relating to the individual topics covered.

In analyzing, summarizing, and organizing the literature and in compiling this review, extensive use has been made of the terminology and mode of expression of the authors of the articles cited. Such passages are quoted, with references to their sources, but with sparing use of quotation marks to make for easier reading of the text. The authors' own words were used in this way to ensure clarity and to avoid possible misinterpretation.

II

SYSTEMS FOR STORING AND SEARCHING CHEMICAL STRUCTURES

*

A. PURPOSES AND USES OF SYSTEMS FOR SEARCHING CHEMICAL STRUCTURES

The fundamental method of depicting organic chemical structures in documents is by structural diagrams, or, as has been said, the ". . . common denominator to the heterogeneous technical knowledge that must be managed efficiently is chemical structure."[13] Such diagrams convey for a given specific substance (a) what elements are present, (b) which atoms are connected to each other in a molecule of the substance, and (c) by what types of bonds they are connected. For purposes of information retrieval, chemical structures can be represented by graphical diagrams, nomenclature, fragmentation methods, or methods of complete representation.

The latter can be further distinguished, especially in connection with the effective computer-based management of chemical structure information, as two-dimensional arrays of characters representing structures, as do the usual structure diagrams (e.g., the machine stores a set of characters and their coordinates), atom-by-atom and bond-by-bond listings of the parts of the molecule (either in list form or as a set of matrices), and notation schemes or linear ciphers (one-dimensional arrays of symbols that represent the topology of the molecules).

Fragmentation methods are based on the principle that various atoms and groupings of atoms occur in molecules and that concurrences of such groupings are of importance in retrieving the struc-

ture. Two limitations on these methods are that multiple occurrences are not noted in most fragmentation schemes and that there is usually no indication of which fragments are bonded to which or of what atom of a fragment is used for attachment. "A fragmentation method of representing structures is [in effect] a coordinate indexing scheme in which the structure fragments . . . are the indexing terms."[185] The terms are chosen arbitrarily and in advance.

There are a number of operating information systems based on fragmentation schemes, and their continued use is evidence of their value. But what is considered a satisfactory scheme for representing structures is strongly dependent on the requirements put on the system. What works well for recovering information from a relatively small file covering a restricted field of chemistry might not be adequate for, or even adaptable to, retrieving information from a file approaching universal coverage, such as that of the U.S. Patent Office or that of the Chemical Abstracts Service (CAS).

Complete representations can include notations or ciphers, connectivity tables, topological coding schemes, and related devices. The fundamental advantage of a mode of complete representation is that the original structure diagram can be reconstructed because no information has been lost. "Each of the unambiguous representations is isosemantic to the structure, and, of necessity, each is isosemantic to each of the other unambiguous representations of the same structure."[185] By any of the systems of complete representation, the chemical structure can be presented in a machine-usable form, at least to the extent that the systems have been worked out. The use of one such system rather than another is, as suggested earlier, a matter of expediency, personal preference, and the requirements put on the operating system.

This report reviews developments in structure-searching systems through the spectrum of fragment codes, linear notations, and topological and graphical coding systems,[87] ranging from systems easily applied and familiar to chemists, through systems requiring extensive rules but still recognizable to chemists, to systems depending on manipulations within a computer to describe structural features.

B. FRAGMENTATION TECHNIQUES

As defined in Survey of Chemical Notation Systems, fragmentation codes assign code designations or terms mainly to specific groups of atoms and bonds. The results are ambiguous codes that are not completely descriptive of structures. As explained in that report,

fragmentation-type codes were applied early for correlation studies, and several systems then in operation were described in detail. Some have undergone further development and are reviewed here along with systems designed and put into operation since that report was written.

National Cancer Institute

The Chemical Information Retrieval System at the Cancer Chemotherapy National Service Center (CCNSC) of the National Cancer Institute was established for a broad analog searching of chemical structure.[91] It is based on the National Bureau of Standards Peek-a-Boo System, which uses cards in the form of 5 × 8 inch, 8–10 mil vinyl plastic sheets. One or several descriptors are used for selection by overlaying the plastic cards; the search can be varied if the desired information is not found or, conversely, if too much "chaff" appears.

Changes that have been introduced in the system of coding chemical structures involve the ring systems and the fragment chains. In earlier sets of descriptor cards, only aromatic rings, plus a few special heterocyclic rings, were recorded as units. Partially hydrogenated aromatic rings were split into fragments. In the present system, all fused and spiro rings are kept intact as indexing units. Each ring system in the Ring Index has a serial number that is coded into the Peek-a-Boo cards. A special group of 44 cards is numbered for thousands, hundreds, tens, units places (10 cards each), and the tenths place (0.1–0.4). A four-place serial number, the Revised Ring Index Number (RRI#), is punched into the four appropriate cards. It is stated that more than 10,000 ring systems can be recorded and retrieved with these cards.[91] Occasionally, two or more ring systems may occur in the same structure. About 22 of the most common and frequently occurring rings have been entered on separate cards, partly for easy retrieval and partly to reduce the need for punching two numbers for the same compound, with false readings of numbers resulting during retrieval. To retrieve ring systems in a more general way, descriptors may be used. The prefix "hetero" refers to substitutions in rings of elements other than carbon; several hetero number cards may be punched, depending on how many ring systems there are in one compound. The new method for coding ring systems has been found to be very specific when using the RRI# and quite general when using the allied descriptors, so searching time is lessened and fewer false drops occur.

The fragment chain is defined as a series of atoms, linked one to

the other by single or multiple bonds in which there are no singly bonded carbon-to-carbon atoms. Most organic radicals and functional groups attached to carbon chains or aromatic rings are indexed as units if there are less than four atoms in the radical or group. The revision introduced in the treatment of fragment chains involves groups having four atoms in a chain or those having three atoms in a chain plus one or more branches on the chain. A special group of sixty plastic cards is used to code the large number of combinations possible with nitrogen, oxygen, phosphorus, sulfur, and carbon in any one of five positions in a chain. An order of preference has been established for coding these chains. Over 300 different fragment chains have occurred in the first 5,000 organic compounds received in one year (1964). A fragment chain may be wholly or partially included in a heterocyclic nonaromatic ring, in which case it is coded both as a fragment chain and by its RRI#.

Abbott Laboratories

The punched-card file of chemical and biological data at Abbott Laboratories, described in Survey of Chemical Notation Systems, is being converted to a tape-oriented computer system.[134] A fragmentation system is used for indexing and retrieving chemical structures. A complete interpretation of the structures and information for the investigator was desired—a printed record that would remain up-to-date as new data data were accumulated and that would be in a language familiar to the investigator. The organic chemical fragmentation code describes chemical compounds according to ring systems, carbon chains, and functional groups present. The code is a multiple-punch code representing structures in tab cards for machine searching. Structures are also hand-drawn on the tab cards into which they are coded. This multiple-punch code was converted from the earlier electronic accounting machine (EAM) system in stepwise fashion to the tape record for computer searching.

Organic chemicals are most commonly expressed and understood in terms of two-dimensional structural formulas. Therefore, the HECSAGON chemical structure display proposed by Horowitz and Crane at Eastman Kodak Co. is used for computer printout of chemical structure. (This system is discussed in a later section of this review.) The structure of each compound is drawn on quadrille-ruled paper, edited, and key-punched, using one card per line of structure. One person working about three days per week drew some 20,000 HECSAGON structures in seven months. A computer program was written to compact the HECSAGON structures onto magnetic tape

and print them out from the tape with the Burroughs 280 computer. In this way the investigator will receive a printed structure of each compound that meets the requirements of his search request.

Still another card file had to be generated for input to the tape record, giving for each compound the molecular formula, the source, the notation of any clinical work, and the date of coding. This card is used with a computer program to compute the molecular weight for each compound, and a new card is then generated containing the preceding information as well as the molecular weight. This is the first card read onto the tape record of a compound. Biological data are contained in two major card files, one containing toxicity and symptomatology data and the other containing screening data. The computer program to interpret and print these data is a very significant asset to the automated system. The tape record of a compound, then, contains the following information: Abbott number, molecular weight, source of the compound, notation of any clinical work, accession date, molecular formula, two-dimensional HECSAGON structural formula, chemical-structure search code, toxicity and symptomatology information, and data from screening programs. Any part of the record can be searched and printed out as desired, with the exception of the HECSAGON formula, which is used for printouts only. In a search, the "hits" are written onto a hit tape, which is then used with the print–edit programs to interpret and arrange the output in any order desired. The hit tape is retained until the user is satisfied that the output has been arranged in as many classifications as he needs.

Kodak Research Laboratories

Kodak Research Laboratories at the time of the Survey of Chemical Notation Systems had a system under development to index and retrieve chemical information from a large collection using tabulating cards and standard tabulating equipment. Provisions have since been made for storage and searching using computers.[79]

Various features of chemical structure are separated into several major groupings. Each major grouping is assigned to a separate card, and specific columns in that card are assigned to the various functional groups that may be a part of the major group. For example, oxygen functions outside rings are recorded on one card, and specific columns are reserved for specific groups such as carbon–oxygen, oxygen–hydrogen, or carbon–oxygen–hydrogen. The cards are filed in order of the major groups and can be divided into subdecks by specific function. The column reserved for one functional

group can also show the manner in which the functional group is connected to the other atoms in the molecule and the number of times a functional group connected in a particular manner is present in the molecule.

It is claimed that many thousands of compounds can be placed in the system before it becomes too unwieldy for sorter-collator searching so that conversion to computer operations is required. However, no report of operating experience with this system has been found.

Rotadex and RotaForm

Another system under development at the time of the Survey of Chemical Notation Systems is the Rotadex (Rotated Index) of the Institute for Scientific Information (ISI). "The proposed system is a multiple listing of the molecular formulas and generic structural codes for chemical compounds."[52] The four-place structural codes to be used in Rotadex are generated by looking up key aspects in four tables of chemical features. The assignment of chemical features to these tables is tentative. As an example, N, N-dibutylacrylamide could be coded 99 QT; the first 9 describes the absence of any homocyclic ring structure or spiro configuration, the next 9 indicates the absence of any heterocyclic ring structure or bridge configuration, Q shows that only one of five chemical configurations of oxygen or sulfur is present, and T indicates a tertiary nitrogen and a double or triple bond. The addition of a fifth character to the codes might be useful if it were desirable to indicate additional chemical features such as the presence of metallic salts, organic salts, polymers, or incompletely known structures. With four characters, one may write 32^4, or over a million, different structural codes. Though almost every one of these codes is theoretically acceptable, they will not, of course, find equal use in actual applications. From a study of use frequencies will come improved assignment of chemical features to the tables.

The Rotadex structural codes were designed to facilitate the unambiguous assignment of compounds to proper categories. Generic searching with Rotadex is facilitated by a fixed-column format. The rotation of molecular formulas and chemical codes is performed without actually moving the characters out of their fixed-column positions. Though proposed as a system to facilitate hand searches of printed indexes, Rotadex would seem to be suited to mechanical and computer searches also. No report of operating experience with the system has been found.

Another proposal from ISI suggests a rotated formula index, RotaForm Index, prepared as a by-product of the molecular formula indexes to Index Chemicus. The computer duplicates the formula as many times as there are different elements. A search for compounds containing a specific element is facilitated because all the elements are sorted alphabetically. The RotaForm Index supplements the conventional formula index, permitting a limited range of generic searches, and is particularly valuable for a search involving the less frequently occurring elements. This index "simply advises the reader which molecular formulas contain a particular combination of elements."[66] As a hypothetical example, C23 H20 Al2 Br3 F4 Na2 P3 would be repeated under the following arrangements:

Al2Br3F4Na2P3C23H20
Br3Al2F4Na2P3C23H20
F4Al2Br3Na2P3C23H20
Na2Al2Br3F4P3C23H20
P3Al2Br3F4Na2C23H20

Again, no report of operating experience for this suggested tool has been found.

University of Dayton

A fragmentation scheme employed by the University of Dayton is based on the concepts of basic structure, substituents, and connectors, which make possible the handling of a large number of organic compounds by means of a relatively small number of fragment terms.[99] Basic structures are named according to the International Union of Pure and Applied Chemistry (IUPAC) nomenclature, except for widely used common names; in more complicated compounds, the basic structure is determined by the most important functional group.

Atoms, groups, or radicals that usually replace a hydrogen atom in an organic compound are used in the system as substituents. Simple substituents retain their identity, but more complicated ones are fragmented. Claims made for the system are (1) its ability to reconstruct any organic compound from its fragments, (2) its ability to search for a specific organic or organometallic compound, and (3) its ability to search for a class of compounds. This capability is effected by the use of a special device called "U.D. Connectors," which are sets of two consecutive digits indicating the connection of a substituent with a basic structure.

This fragmentation system has been applied to organometallic compounds, metal complexes, and organic compounds of boron, phosphorus, and silicon. The system is designed to accommodate a large number of organic compounds by means of a relatively small number of fragment terms. As a consequence, there is some degree of uncertainty, and screening of the search output is required. But index and abstract cards provide the necessary information, so the screening does not require reference to the original documents. Materials are associated with properties and processes by means of links sufficient to designate the correlation of an organic compound with a physical process, a property, or a use.

The system has been extended to an indexing vocabulary and coding system for high polymers.[100] The vocabulary is based on available nomenclature schemes, such as those of IUPAC and Chemical Abstracts (CA), as well as on common or trivial names.

Imperial Chemical Industries, Ltd. (ICI)

In the Pharmaceuticals Division of ICI, the research department has synthesized and examined for physiological activity over 50,000 different organic compounds submitted for nearly half a million separate biological examinations covering the fields of both human and animal diseases. "To provide rapid access to this vast amount of information, the technical relations and intelligence section has installed an International Computers and Tabulators Ltd. (ICT) punched card system, including a system designed for the representation of chemical structure on punch cards."[74] This system was discussed briefly in Survey of European Non-Conventional Chemical Notation Systems.

The code used is basically a fragment code showing some linkage between the fragments. Altogether there are 288 coding features, and the majority of compounds are represented by between twelve and twenty of these features. The chemical code occupies twenty-six columns of an 80-column card; the remaining portion of the card is allocated to availability of sample and biological test data. By the last quarter of 1962, when the collection approached 50,000 compounds (at a growth rate of about 5,000 compounds per year) and searches totaled 25 a month, the saturation point had been reached for the system then in operation. There was also a gradual change in the type of question asked. Questions were increasingly directed toward searching for structures in which atoms occur in specific order, requiring the machine to search for numerous alternate patterns in one card passage.

As a result of these demands, the newly developed ICT 335 statistical sorter, linked to a summary gang punch, was installed. This machine has an increased capacity for pattern selecting, so searches can be grouped. It also has the ability to extract information from the card and to reproduce it in punched-card form by way of the summary gang punch, thus preserving the sequence of the master file.

A new service available to the biologists and chemists is the analysis of their work on a particular disease at suitable intervals. Often these analyses cover only one or two chemical classes, but sometimes, for statistical purposes, it is necessary to give full cover of the entire file of compounds for which testing and activities have been recorded. Such information calls for counting the frequencies with which fragments of molecules occur in both the active and the inactive compounds and presenting the results in the form of probability tables for further study by the research team. Master cards contain chemical code, reference number, details of all biological tests performed, and availability of samples. Detail cards contain new biological and sample data to be transferred to master summary cards. Catalogs, some in the form of dual dictionaries, have been compiled. They provide guidance to the structure of compounds, each of which is given a serial number. Some of the questions require the structures to be searched atom-by-atom; the present system allows for conversion to computer operations, via punched cards or tape, whenever necessary.

Another system, based on numerical codes, was developed at the Fine Chemical Service of ICI.[33] The system was operating with punched cards, but it was decided to program typical searches experimentally for a Pegasus computer. The technique for coding the file items was based on prime numbers, which resulted in savings in storage space.

In the punched-card system, different structural features were identified (for example, ring systems, amines, chlorides), and fixed positions on a card were allocated, one to each feature. By analogy, bits could have been set aside for recording structural features in the tape system and, if a compound contained a particular feature, the corresponding bit would have been made unity. But only a small fraction of the possible configuration of bits would have been used, resulting in inefficient utilization of storage space. Therefore the first 208 prime numbers were allocated to the 208 structural features. For any chemical compound, the primes corresponding to its structural features were selected and multiplied, and the resulting product was used as the code number for the chemical compound. The presence of a particular set of structural features is tested by dividing the code number by a factor formed from the primes for the

corresponding features and testing for zero remainder. The programming is no more complicated than the alternative of using collate and not-equivalent operations followed by a test for zero result. No report of operating experience with this system has been found. More recent system developments at ICI are reviewed in later sections of this report.

BATCH Number Fragmentation Codes

As early as 1953, Wiswesser had proposed for organic compounds a Formula Index Number consisting of five simple numeric measures, each having possible values from zero through nine.[97] This structural "atomic" code was given the mnemonic designation BATCH Number. In 1964, the definitions for the assignment of B and A digits were revised on the basis of statistical study; the current assignments are: the B digit relates to the nature (and number) of rings in the structure; the A digit relates to the nature (and number) of atoms present other than carbon, hydrogen, and oxygen; the T digit indicates the number of atoms present other than carbon and hydrogen; and the C and H digits are based on the carbon and hydrogen atom counts.

"The BATCH Number has been demonstrated to be of value in the organization of structural-atomic indexes, which have been termed BATCH Directories."[13] Additional search discrimination can be obtained from the BATCH Number by maintaining a single order of the measures but treating various digits as primary, secondary, etc. in sorting to a numeric order. For example, when the BATCH Number is treated as a conventional five-digit number, rings are given precedence over the type and number of atoms present. When the A digit is treated as primary, the T digit as secondary, the B digit as tertiary, etc., types of atoms are given precedence over rings.[14]

The BATCH Numbers also have proved effective in the management of small collections in nonmechanized card files.[12] Color coding facilitates the use of such files.

Since the BATCH Number, as a nonunique code notation, only focuses a search on a small number of entries, some further discrimination must be provided. For storage and retrieval purposes, each chemically distinct functional group or similar structure fragment can be assigned an arbitrary address. These addresses constitute a fragmentation code. In fact, some degree of meaning can be worked into the assigned addresses, and to this extent a fragmentation code becomes a fine-structure elaboration of a general classification number, that is, it becomes a fragmentation number. Fragmentation

codes have long been used in hand, edge-notched card, and punched-card management of organic compound files. Such traditional codes can be entered into a computer, there serving as tools for the answering of search questions and the preparation of listings by such codes.[11]

FMC Corporation

At the Niagara Chemical Division of FMC Corporation, a chemical structure system not covered in the Survey of Chemical Notation Systems is "based upon a combination of empirical formula and functional group classification," and is adaptable to data processing equipment.[206] The functional group classification is based upon the classical approach to organic chemistry; the distinctive features of the system are the handling of oxyacids and the handling of the ring systems.

The various categories of oxyacids having central atoms of valence 2, 4, or 6 are designated on the basis of the corresponding sulfur acids, viz., sulfenic, sulfinic, sulfonic, sulfurous, sulfuric. By combining a search for the appropriate derivative in the oxyacid section with a search for the particular element involved and the analog punch, retrieval of selenium oxyacids and their distinction from the corresponding sulfur acids are relatively simple. In the same manner, the oxyacids having central atoms of valence 3, 5, or 7 are defined on the basis of the corresponding phosphorus acids. Again, multiple searching by combination punches allows distinction between phosphorus acids and other acids having the same central atom valence, such as those of arsenic, antimony, or boron.

The handling of ring systems is a difficult aspect of the coding of organic compounds. The FMC system is based on the total number of rings present in the molecule and their types. A problem arose in attempting to devise means for classifying bicyclic, heterocyclic, and heterobicyclic compounds. In the case of bicyclic or polycyclic ring systems, the basis is the length of the chains joining the bridgehead atoms of the bicyclic structure. The handling of heterocyclic rings is further complicated by the need to distinguish between the various heterocyclic rings in a particular compound. They are classified first on the basis of the number of oxygen, nitrogen, or sulfur atoms in the ring. The heterocyclic rings within the molecule are further classified starting with the largest ring. When two or more rings in a molecule are of the same size, the ring having the greatest number of hetero atoms is given priority, and so on, in descending order.

Crompton and Knowles Corporation

A numerical system of coding organic structures by functional groups has been employed in a chemical information retrieval system at the Crompton and Knowles Corporation.[56] The "code is most useful with small to medium-sized collections of structures," and approximately 2,500 structures have been coded by the method. Each structure is coded as a six-digit number, each digit having a value from zero to nine. These six digits are selected by consideration of: (1) skeletal structure; (2), (3), (4), and (5) location and composition of hetero connectives and functional groups; and (6) carbon count of the structure. The first (left) digit indicates skeletal structure or ring number. Since the company is concerned primarily with textile dye intermediates involving especially high percentages of naphthalene and anthraquinone structures, the use of this digit was modified to give special treatment to these frequently occurring structures. The next (center) four digits indicate directly the presence (to a maximum of four) of hetero atomic connectives and functional groups included within or attached to the basic skeletal structure. These digits and their order in the code are determined by visually scanning the structural formula according to listed search-order phases. Hetero connectives and functional groups found during each of the search phases are cited in the order determined by their established order of precedence. After four digits have been listed, any remaining possibilities are ignored. Search-order phases take precedence over functionality. The last (right) digit of the code number is the units digit of the carbon count in the empirical formula of the compound.

Case Western Reserve Medical Coding Scheme

A chemical coding scheme developed for use in the Comparative Systems Laboratory (CSL) of the Center for Documentation and Communication Research, Case Western Reserve University, is contained within the Medical Coding Scheme (MCS) developed there. The end product of MCS is a faceted-type thesaurus in symbolic language suitable for computer handling, which arranges index terms by employing individual classification schemes for various subject groups of terms.[18]

One of the categories within MCS is the chemical coding scheme, which provides a classification scheme and encoding method for drugs and chemical terms. The chemical code is based on the fragmentation principle, with chemicals classified by the type of units or

groups present. The code does not assign any importance to the sequence in which the functional units appear, and each group is coded individually without reference to other parts of the molecule. Each code may be thought of as an entry in a faceted-type thesaurus in which chemicals sharing certain characteristics are grouped together, facilitating cross-searching. The thesaurus-like relationships of the code allow searching on an extensive basis.

The chemical divisions are: (1) inorganic, dealing with all chemical elements and their compounds; (2) organic, carbon-containing substances, subdivided into aliphatic and alicyclic, and aromatic and heterocyclic; and (3) organometallic, compounds having carbon–metal bonds, the word bond being taken to include all types of chemical combinations or linkages that do not involve an intermediate atom. The CSL dictionaries include approximately 11,000 terms, 2,500 of which are chemical terms employing the chemical coding scheme.

According to the authors,[18] the chemical coding scheme, like other systems, has faults and limitations. Among the most important, as shown by use, are the following: (1) the codes were limited by the system to eight punches on the IBM card, (2) not all compounds belonging to a particular structural system will have the same general chemical classification, (3) there were difficulties in setting up mutually exclusive classes and in arranging the order in which characteristics or division should be applied, and (4) no generally satisfactory way exists for classifying organic compounds as to aliphatic, aromatic, or heterocyclic character.

Gordon Hyde Science Communications

The London firm of Gordon Hyde Science Communications has proposed a coding system for chemical compounds to be used in documentation retrieval. The compounds are identified by a seven-digit code in place of combinations of descriptors.[34] The system is said to be superior to word-indexed systems in specificity, in storage capacity, and in simplicity of programming for updating and retrieval. It is compatible with manual, microfilm, punched-card, or computer retrieval and can be used as a machine language between different types of retrieval. With a total capacity of more than 300,000 descriptors, some 700 of which are used in the chemical code, the system is claimed to have the capability of covering every field of chemical technology and research with addition of suitable subject descriptors.

GREMAS System

Generic Retrieval by Magnetic Tape Storage (GREMAS) was described in Survey of European Non-Conventional Chemical Notation Systems as "widely useful and worthy of further trials." Further experience with this faceted classification system in organic chemistry was described at the Elsinore (Denmark) Conference on Classification Research.[64]

JICST Code

The Japan Information Center of Science and Technology (JICST) has been conducting, since 1963, retrieval experiments using a fragmentation code, linear cipher, and atom-by-atom topological code. The advantages and disadvantages of the respective systems have been examined.[189] The results of the comparison were tabulated as Table 1 illustrates.

JICST concluded that a fragmentation code seems to satisfy its requirements and is now developing its own system. In looking to future developments, the center concludes that even if an atom-by-atom topological system is to be implemented eventually, a fragmentation code system will still be necessary for screening purposes.

TABLE 1 JICST Comparison of Systems[a]

	Fragmentation Code System	Linear Notation System	Atom-by-Atom Topological System
Encoding Rule	Simple, and easy to understand, but needs chemical knowledge.	Voluminous rules, taking much time to understand. Chemical knowledge needed.	Rules are the simplest of all three, and anyone can encode.
Code List	List needed, but far less in volume than old systems.	List unnecessary.	List unnecessary.
Encoding Speed	Average of 1 or 2 minutes per chemical compound.	Average of 5 minutes per chemical compound.	Average of 15 minutes per chemical compound.
Input Data	60 characters per chemical compound.		477 characters per chemical compound.

[a]The particular systems were not further identified in this reference.

Silk Fragment Code

At the meeting of the American Chemical Society (ACS) in San Francisco, April 1968, Silk described a notation-based fragment code.[162] The notation is used to construct meaningful code terms for molecular fragments, thus providing greater flexibility and specificity than a system based on a predetermined set of fragments. The code has been applied to pesticide patents and currently provides the means of searching nearly 20,000 patents to retrieve comparisons for purposes of analysis and correlation.

The chemical code was derived mainly from the original Silk notation described and referenced in Survey of Chemical Notation Systems and in Survey of European Non-Conventional Chemical Notation Systems. Considerable modification was needed to make the system suitable for fixed-field coding on punched cards. The card system has been supplemented by a computer-based technique that provides comprehensive classified lists of patents. The format of these lists, consisting of standardized one-line notation entries, makes them easy to scan. Many queries can be answered by inspecting these lists instead of searching the file. The two systems thus complement each other usefully.

C. LINEAR NOTATION SYSTEMS

Notation systems offer a concise representation of structure and a useful tool for indexing chemical literature. As previously mentioned, two of the possible means for the retrieval of chemical structures are a selective, nonunique fragment code that permits retrieval of a group of structures, and a unique notation that allows unambiguous retrieval of a single structure. In Survey of Chemical Notation Systems, the phrase "chemical notation" was used for any representation of a chemical compound, and more specifically for non-conventional representations. There the term "notation" was equated with the phrase "chemical code," further characterized as classification code, fragmentation code, topological code, or unique, unambiguous notation completely descriptive of structure. It is in the latter sense that the phrase "linear notation system" (usually shortened to "notation system") is used in this review, that is, as a system for the unique and unambiguous linear representation of chemical structure.

The fragment codes discussed in the previous section offer the advantages of a system that can be mastered by the user in a short time and that even provides him with a classified directory of struc-

tures. Notation systems, discussed in the present section, also offer some ease of use. "A chemist can write notations for relatively complex structures more readily than he can write correct systematic names."[19]

If a notation is used for the construction of a manually operated type of index, such as a card file or a list, the principles used in designing the notation will control the type of organization of structures that will result from alphabetizing the notation. The system of rules will put certain features of structures into indexing prominence and will subordinate other features that will then be scattered throughout the index. Consequently, the index will perform poorly for manual use if the subordinated features are the items in which the searcher is interested. However, machine manipulation of notations can overcome this type of problem, and the systems to be discussed next have developed such machine techniques.

The design of a linear notation system and the clarity of the system's presentation have a marked effect on the ease and accuracy of its use. It is extremely difficult to design rules to handle all foreseeable situations and to state them with precision and clarity. This difficulty is reflected in coding accuracy. Any individual coding at a reasonable rate of speed and not going back over his work is going to have some significant level of error. Even after independent checking of the coding effort, residual error of some magnitude probably remains.

Notations, to be successful, must operate with a set of logical rules; but inevitably, changes will be made in notation rules, which may necessitate extensive revisions in notation indexes. Computer programming, however, can make file changes practical with surprisingly little manual effort.

The sheer magnitude of the problem of indexing chemical information will force changes from past practice. The use of notations could significantly reduce the amount of cross-indexing required. Consequently, newer techniques have been, or are being, tested and evaluated in the effort to develop even more effective tools for searching structures.

The two ciphers or notation systems of major interest over the past several years are the Wiswesser Line Notation and the IUPAC (Dyson) notation system. The former has been applied broadly in a number of installations. The latter has had limited use. Several other systems have been developed for specialized operations.

Wiswesser Line Notation (WLN)

The Survey of Chemical Notation Systems listed five organizations and individuals using the Wiswesser notation in active or experi-

mental information-retrieval systems. As of June 1968, the list had grown to thirteen, at least three of which have encoded over 100,000 structures in Wiswesser notation.

Development has taken place in the four main areas discussed below.[166]

1. Permuted notation indexes Studies have been made of the feasibility of permuted indexes of Wiswesser notations and of methods for producing them by means of tabulating equipment for small-scale files and computers for large-scale files. (These studies are reported in more detail later in this section.) Programs exist for making permuted WLN indexes on UNIVAC, IBM, Honeywell, Burroughs, and GE computers.

Such indexes serve very well to locate ring systems, functional groups, and structures that are expressed explicitly in the notation, but they are not convenient for locating structures that are only implicit in the notation. Searching for the latter structures would require computer processing of the notation itself.

2. Algorithms and computer manipulations Exploratory and experimental computer programs for the manipulation of Wiswesser notations have developed in several directions: (a) machine checking of correctness of encoding, (b) machine conversion from card input to magnetic-tape registry files and information files, (c) machine fragmentation of notations to give an inverted searching file, (d) machine searching of files, (e) machine printout of recognizable structural formulas from notation input, and (f) machine production of connectivity tables from the notations.

3. Organization of users of the Wiswesser notation In the spring and fall of 1964, a group of nine chemists met to study and vote on proposals for changes in the Wiswesser notation rules. The nine agreed to form an organization, The Chemical Notation Association, to act as an official body for standardizing and controlling the rules of the notation and to further its development in other ways. There was need, for example, to specify the procedures for controlling the rules of the notation, which, of course, must grow and change with the growth in knowledge about chemical structures. The group included individuals with extensive encoding experience who had access to sufficiently large files of notations to enable them to test proposed rule changes experimentally, and who wished to share the responsibility of controlling the development of the notation on the basis of experimentally acquired evidence.[167]

4. Revision of the Wiswesser manual The revision of the Wiswesser manual, including the latest changes considered, was carried out under the guidance and control of The Chemical Notation

and reflects the accumulated experiences of some thirty chemists in encoding several hundred thousand structures. This revised manual, published in 1968, describes in detail the Wiswesser chemical line notation system.[167] It gives complete encoding and decoding instructions for most classes of compounds.

Times required for learning the Wiswesser notation from versions of the revised manuscript were recorded by The Chemical Notation Association.[166] They found, in one instance, that a beginner can start encoding compounds, at least the simpler ones, within a few days to about two weeks. Within three months, coders can be fairly proficient if they are doing a considerable amount of encoding. In another example, it was found that a chemist learns the notation in two weeks. During this time he reads the manual, handles problems, encodes structures coming into the system, and reviews his errors after the notations are proofread. After two weeks, very little supervision is necessary because all proofread notations are routinely returned to him for inspection of errors. Subsequent coding practice seems to increase speed markedly but does not greatly affect error rate.

Members of the association state that they consider the notation to be: (1) the easiest, cheapest, fastest method of preparing computer input of this type; (2) the most concise representation of structure they know; (3) capable of serving as its own registry-file device; (4) a language convertible by computer program to many other forms; (5) a language useable from the lowest level of mechanization through hand-operated edge-notch cards and other simple systems, to punched-card systems, to small computer systems, to the largest computer installations; and (6) an indexing device for chemically oriented literature that is capable of offering an almost perfect list of unit terms or descriptors. One does not need a thesaurus, there are no synonyms or homonyms, there is no redundancy in meaning, and great conciseness is the rule.[166]

The Wiswesser Line Notation represents an important advance in the efficient machine handling of information about chemical structures. Granito and his co-workers suggest that the notations are intelligible at sight to any chemist who studies them, so it is possible to employ computer preparation of chemical structure indexes with which many generic structure searches can be made without additional use of the computer.[70]

The system uses a blank space as a symbol that breaks up the notation into small, word-like symbol groups that are easy to read. Letter symbols are used to denote functional groups, and numbers are used to express the lengths of alkyl chains and the sizes of rings. These symbols are cited sequentially from one end of the molecule

to the other. All 40 symbols in use (10 numerals, 26 capital letters, 3 punctuation marks, and the blank space) are available on existing computers and punched-card installations, so no modifications are needed.

The line notation system brings together related compounds when their notations are arranged alphabetically in printed lists. However, a slight change in structure could result in a wide separation of notations of similar compounds. A search of such printed lists for compounds possessing the same functional group or atom, e.g., chlorine (represented by G) is obviously impractical. One possible solution reported is a list of permutations of chemical line notations alphabetized on individual symbols.[170] Such lists are used to locate all compounds containing any specified functional group, as well as specific compounds and specific classes of carbocyclic or heterocyclic structures.

Developments in WLN

Most of the recent technical developments in the Wiswesser Line Notation have been reported in three groups of publications. The earliest of these, which appeared in 1964–1965, were from the Industrial Liaison Office of the Army Chemical Research and Development Laboratories at Edgewood Arsenal, Maryland. They described the technique for generating permuted notation indexes that make possible many generic searches that are not feasible in ordinary alphabetized indexes. The second group comprises the Proceedings of the Wiswesser Line Notation Meeting of the Army Chemical Information and Data Systems Program in the fall of 1966.[132] These papers discuss use of the notation for registration, storage, and retrieval of structures in several industrial and governmental collections; generation of structural fragment codes by computer from the notation; a partial algorithm for development of connection tables from the notation; and many details regarding costs and techniques for encoding and searching. The third group of papers, for the most part presented at national meetings of the American Chemical Society in Miami (spring 1967) and San Francisco (spring 1968), deals with computer programs for manipulating the notation in various ways. The most important of these are (1) a program that checks accuracy of encoding by calculating the molecular formula from the notation and comparing it against the known molecular formula; (2) programs that convert the notation to a connectivity matrix and search the matrix for substructures; (3) a program that generates the canonical notations for complex polycyclic structures from sim-

ple ring-connection input data; and (4) a program that generates displays of structural formulas from notation input. Each of these groups of papers is discussed in turn.

In 1964–1965, Sorter, Granito, Gelberg and others at Edgewood Arsenal described a means of preparing the permuted index of Wiswesser notations that counteracts the scattering of related structures produced by ordinary alphabetized indexes.[170] The method was originally described for use with a card punch, a sorter, and a printer[75]; later it was adapted to various computers and programmed for the UNIVAC II, the IBM 1401, and the Honeywell 400 computers, among others.[76] In this procedure each notation is scanned for the presence of notation symbols on which it is desired to index the compound. When such a symbol is found, the notation is shifted to the right or left in the printing field to bring this symbol into a central vertical index column. The notation is repeated in the index as many times as necessary to index all its pertinent symbols. The notation list is then alphabetized on those symbols lying in and to the right of the central vertical indexing column. The resulting permuted list of notations is printed out, making possible manual searches for all compounds having any substructure explicitly stated with the notation symbols indexed.

For the 55,000 compounds in the Edgewood files at that time, an index containing 350,000 notation entries was generated, a cross-reference ratio of about 6.4 notation entries per compound. Estimated costs of about $12,000, or $0.22 per compound, are stated,* including all encoding, card punching, verification, programming, machine time, and other direct costs.[76] Additional compounds are added from time to time by a program that updates the computer tapes. By 1968 these indexes contained more than 100,000 compounds, and they have met satisfactorily the structure searching needs of this organization.

Since the Edgewood compounds are of a proprietary nature, these investigators also produced a permuted notation index from public information sources, both for their own use and for public demonstration of the system.[210] The catalogs used included: (1) Gelberg's checklist of the structures in Frear's Pesticide Index, 2nd edition; (2) Sorter's checklist of the organic structures in the Merck Index; (3) all the organophosphorus pesticides in Eugene E. Kenaga's 1966 revision of Commercial and Experimental Organic Pesticides, which excluded those no longer available experimentally or commercially; (4) most of the organophosphorus compounds in the Chemical-Biological Coordination Center accessions that had been screened at

*Cost estimates should be viewed with caution, as noted in the Preface.

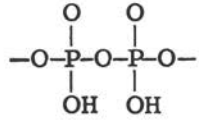
USDA's Beltsville (Maryland) Center and at Fort Detrick, Maryland; and (5) other old and unclassified organophosphorus compounds that had been screened at Fort Detrick. The sets of cards for these catalogs were put through an editing program at Edgewood Arsenal to eliminate duplicates before the notations were permuted. Only exact matches were recognized as duplicates, so slight variations in coding were accepted as nonduplicates (the encoding rules were being changed at that time).

It is instructive to see how structural correlations among organophosphorus pesticides are revealed by this permuted notation list. The seventh page of the permuted listing shows the beginning of those notations containing letters, "A" for unspecified alkyl groups, "C" for "carbyl" carbon atom in nitriles, isonitriles, and the like, or two-letter symbols for metals such as "CA" set off in hyphens, "E" for bromine atoms, and "G" for chlorine atoms, with "F" for fluorine falling appropriately between them. The next six pages of the permuted notation list reflect the abundance of chlorine atoms in these compounds.

Additional combinations are recognized in terms of the corresponding functional-group combinations because the elementary atomic definitions of the alphabetic symbols have been kept intact. Thus, long strings of letters such as . . . OPQO&OPQO&O . . . are comprehended just as quickly as any chemist scans the corresponding line formula:

OP(OH)(O)OP(OH)(O)O

or diagram:



The notations are, in fact, nothing more than line formulas, standardized in terms of the limited set of symbols available in high-speed printers to enable such computer-generated correlations to be produced.

The second group of papers, appearing in the volume covering the Wiswesser Line Notation Meeting of 1966,[132] contains a wealth of detail and experiences. Bonnett[21] discussed the classified notation indexes used at G. D. Searle and Co., which contain some 125,000 compounds, and he reconfirmed the reported input cost of $0.25 per compound for generation of the printed indexes, starting from the drawn structural formula. Horner[85] had rewritten the Edgewood permutation program in extended ALGOL-60 for a Burroughs B-5500 computer. He used this program to prepare two permuted indexes of

small files, one for the 2,800 compounds in the files of Stanford Research Institute, and the other for the 1,850 compounds in the four collective volumes of Organic Syntheses. The total direct costs for the two indexes were reported to be about $0.19 per compound, including cost of programming, encoding, card punching, verification, and computer time.

Granito[77] reported on the use of the computer to update the Edgewood notation files. A registration run for 7,000 new compounds as input to a file of 86,000 notations required 90 minutes of computer time, including card-to-tape input, sorting, checking, registration, updating, and printouts, at a cost of about $0.20 per compound. Further details of this process were added by Renard,[153] while Sorter[171] gave details, including a flow chart, of the computer updating of the Hoffmann-La Roche notation files.

Bowman and others[24] described in some detail the Dow Chemical Company's computer program (for a Burroughs B-5500 computer) for generating from the notation and molecular formula a structure-fragment code with more than 40,000 terms. The code is overlapping and redundant to ensure complete retrieval, and it is searched in an inverted file order by a master search program, which is also used for other purposes. This system, as well as others reported at the meeting, is discussed further in the section of this report devoted to interconversion of structure representations.

Munz reported on the formal analysis of notation systems, which is strongly interconnected with the pragmatic approach evident among users.[137] Because there are regular semantic relations between the expressions of linear chemical notations and chemical structures, notations are practical tools for storing and processing chemical information.

A paper by Hyde[88] described an open-ended fragment code generated by computer from the Wiswesser notation. One by Fraction and others[61,62] described in considerable detail the generation by computer of connection tables of the conventional atom-and-bond type for acyclic and benzene compounds from Wiswesser notations. In general, this one volume gives a good idea of the state of developments in the Wiswesser Line Notation as of fall 1966.

The third group of papers covers more recent developments and deals with computer programs actually in operation, as distinguished from some of those described earlier. For example, Bowman and others[23] discuss the existence of a computer program at Dow that checks the accuracy of encoding a notation by calculating the molecular formula and comparing it with the correct molecular formula in the input data. A few checks were made on the syntax of the notation as well, but this aspect of the checking program had not been fully

worked out at the time of publication. Among other things, the program calculates the hydrogen atom count, which was found to be one of the most powerful checks of all in determining the accuracy of the encoding. To avoid problems with methyl contractions in the notation, the program looks for those and expands them into an uncontracted form at an early stage. An algorithm was developed for calculating from the notation the number of ring atoms in cyclic structures, and this has proved to be a powerful checking tool, both in the computer program and in manual checking by the encoder.

Bowman and others[25] have also written a program in extended ALGOL-60 for the Burroughs B-5500 that can generate the canonical Wiswesser notation for complex polycyclic structures from simple input statements of the nonconsecutive locant pairs of any continuous path of locants through the ring structure. From this information the computer constructs all the possible paths and compares them with the first seven hierarchic requirements of the fundamental Wiswesser rule for determining the canonical path. A three-ring system with two bridge atoms and one multicyclic point required 8 seconds of computer time for analysis; a four-ring system with two bridge atoms and two multicyclic points, and two internal spiro points, required 124 seconds. Thus the computer time required is not trivial, and the program will be used only for the most complex ring structures where manual encoding is most difficult. The program eliminates the main source of encoding error—the uncertainty that sufficient locant paths have been investigated by the manual encoder to find the canonical one.

Granito described a computer program written to include biological data in an index of permuted Wiswesser Line Notations.* The approach is from the chemical point of view: questions deal first with chemical structures and then with data. Ofer described a computer program written in FORTRAN II that can be run on almost any computer and that can be used to index or search files of linear notations.[142] Although the program was designed to process WLN entries, it can be adapted to other linear notations, according to the author.

Other papers in this third group describe computer programs for converting Wiswesser notations to connectivity matrices for organic compounds. These papers are covered in the section of this report devoted to interconversion of notations.

At the national meeting of the American Chemical Society in San Francisco in April 1968, a tutorial session was held at which princi-

*Paper presented at 155th meeting, American Chemical Society, Division of Chemical Literature, San Francisco, 1968.

ples and techniques for use of the Wiswesser Line Notation were presented.[168] More than 50 persons attended the session to try encoding and decoding the notations. Also at the San Francisco meeting, the Institute for Scientific Information described plans for the Index Chemicus Registry System, using Wiswesser Line Notations.[92]

In summary, these papers and activities indicate the growing interest in the WLN and the adaptability of the system to a variety of machine procedures.

Decoding Study of WLN

At the time Survey of Chemical Notation Systems was published, a decoding study of the Wiswesser Line Notation system was being conducted by the Reading Chemists Club with National Science Foundation sponsorship. The final report, issued in December 1965, is of interest to chemists, although the system itself has been modified.[209] As noted previously, a revised version of the manual for the system was published in 1968; the decoding test was based on the 1954 manual. Nevertheless, the results are applicable to the decoding process with the new manual and the revised system.

The study material consisted of ten notation-reading tests. The ten tests corresponded to the ten major topological groups of chemical structures, sequenced in increasing order of complexity. The decoders numbered some 690 students, teachers, and practicing chemists from about 250 academic, industrial, and other organizations. They devoted a total of 1,015 hours to 2,071 tests, and an additional undocumented number of hours repeating 479 tests on notations they had decoded incorrectly the first time, mainly through oversight or lapses of attention to detail. The high scores of the repeated tests indicated that this method of describing chemical structures is simple and reliable to the initiated. The author concluded that there were no basic decoding difficulties with commonly encountered structure types, other than some minor details already corrected in the revised notations.

IUPAC Dyson System

The Survey of Chemical Notation Systems traced the history of the development of chemical notation schemes and the acceptance by IUPAC in 1961 of the report of the IUPAC Commission on Codification, Ciphering, and Punched Card Techniques. This report took the form of a revision of the Dyson system. The system was examined

in the Survey in connection with the research program of the Chemical Abstracts Service. Some additional developments have been described by Dyson for generic groups in notation programs.[52]

A generic group is one in which the structures composing it have one or more common factors of a specified character. Such groups may be small or very large. An example of a small group is "the dimethyl derivatives of benzene," of which there are three. On the other hand, "the alkyl derivatives of naphthalene" is a very large group. Although useful in many areas of chemical documentation, the concept of generic series and searches is of paramount importance in the field of patents. In chemical patents it is customary to cover groups of related compounds rather than individual compounds. The definition of chemical series for patent purposes is difficult and leads to complex situations. Two main problems arise in this field: first, the means by which a generic search may be made among compounds of definite individual structure, and second, means by which the genericity of a group can be mathematically defined.

The techniques used in performing generic searches on a range of precisely known structures to extract a particular group of closely related structures may be illustrated by selecting structures that are alkyl derivatives of naphthalene. In particular, only hydrocarbons are to be selected in which the naphthalene structure remains intact. The most suitable medium for searching for such substances is a list of notational equivalents of the geometrical structures. The algorithm for locating the required alkyl naphthalenes is based on a search first for $B6_2$, the naphthalene residue. If found, then the letters that follow $B6_2$ must be examined. If $B6_2$ is followed by "C" as the next letter, the search is continued; if followed by any other letter, the structure is rejected. The program goes on to look for $B6_2\Sigma C_{>0}$. Thus, the algorithm pulls from the collection all alkyl derivatives of naphthalene in which all substituents are saturated alkyl groups. Having narrowed the algorithm to the smallest class, it is easy to add modifiers to broaden it in any direction.

The hierarchical system of the IUPAC notation, as published, makes no provision for the ciphering of generic groups, mainly because it was devised for individual compounds. Nevertheless, such groups can be incorporated into the notation. It would then be feasible to use generic notations to express the coverage intent of a patent claim or the allowed extent of such a claim when modified by an examiner.

Additional modifications and provision for using abbreviations in the IUPAC system were described in another paper by Dyson, which was prepared for the ACS meeting in San Francisco.[54] The changes were recommended for computer handling, to decrease the complex-

ity of some of the notations; in particular, the more complex notations are more readily interpreted. Abbreviations are recommended to gain clarity and brevity where basic structures repeat themselves many times as parts of larger molecules.

Shell Research Ltd. System

The IUPAC notation is incorporated into the information storage and retrieval system at the Shell Research Ltd. Woodstock Agricultural Research Center.[42] The research data collection there is small compared with the literature collection, but the technical staff is producing, in reports, in data record cards, and in other forms, ten to twenty million characters of research data per year. The computer storage and retrieval system has been concentrated on these research data.[43]

One of the original justifications for computer use has been the need to search effectively for chemical substructures in a rapidly growing file of chemical compounds. In 1962, it was decided to use the IUPAC notation to convert chemical structures to linear ciphers suitable for computer input and search. The system based on this approach became operational in mid-1965 and covers nearly 50,000 organic chemical structures.[44]

Coding is fast and takes, for an expert coder, an average of about 20 seconds per compound. With less ciphering experience, the coding time is roughly double this value. Keypunching (transferring the cipher to standard 80-column punched cards) is done at virtually the normal speed for an experienced keypuncher, i.e., 3–5 characters per second, or 5–8 seconds for an average cipher. Use of the notation thus enables very fast input of chemical structural information at a low cost. The average cipher length for the 40,000 compounds stored so far on magnetic tape is, including shift signals, 25–30 characters. Storage as variable-length records is obviously of advantage.

The cipher as a search medium has the advantage not only of being compact but also of behaving as an effective screen, the majority of structural features being displayed in the cipher as it stands.[43] It has therefore been economically acceptable to carry out searches by processing the cipher file sequentially. The cost of a computer search has been found to be of the same order as that for the average nonmechanized substructure search using a classified or molecular formula index, particularly if several queries are included in a search run.

The cipher lends itself readily to fragmentation into character-

istic features indicating the presence of, for example, carboxylic groups, a C_5 chain, or keto-groups. Such features are being used to generate a feature-card system for manual (visual) consultation. This subsystem is the most economic and efficient means of dealing with queries in which speed of response is more important than comprehensiveness. The features can also be used to set up suitable file subdivisions in a random-access store. FORTRAN programs have been written to select the features automatically from the ciphers; using the IBM 7094, selection according to a permitted set of 83 different features took about 30 seconds of processor time per 1,000 compounds. Hence, the cost of automatic fragmentation coding would appear to be only a fraction of that for manual coding.

The cipher file can be ordered using conventional sorting procedures. The resulting ordered file provides an efficient means of automatically checking the presence of specific compounds in a large file by matching whole structures (complete ciphers).

The Shell group is now moving toward on-line use of a large central computer installation (UNIVAC 1108 system) backed by very fast drum stores. Marked improvement in search cost and speed of response should result, so it is to be expected that within the next year there will be a sharp increase in the percentage of structure queries answered by using this computer system.

Data on biological properties are keypunched onto punched cards in fixed field arrangement, sorted and tabulated by punched card equipment, and transferred to magnetic tape for computer processing. The system allows selection of compounds according to desired property criteria. The system is under further development. Consultation of the master file will eventually take place through on-line computer facilities.[43]

Swedish Patent Office System

At the Swedish Patent Office, a study was made of the possibility of mechanizing the searching process carried on by a patent examiner.[213] The primary task would be to create a file comprising as much information as necessary for identification of a document. In developing methodology, a file of 60 patent specifications was collected from the class of "Medicines, Organic" (invention resides in the use of a definite organic compound or its preparation for therapeutic purposes). The information stored included chemical composition, i.e., empirical formula, structure formula, Markush variables, IUPAC notation (somewhat simplified), and trivial name or equivalent.

The IUPAC system was simplified to some extent and expanded by including further characteristics. Unknown or variable locants and other numeric characteristics have been symbolized as α. Unknown or variable substituent atoms were represented by β. Markush groups were assigned consecutive numbers; in compounds with variable groups, the corresponding Markush group numbers were substituted for the ciphers for the groups.

The compound ciphers have been split into groups (fragments) in order to institute a screening capacity that might, under certain circumstances, have such a resolving power that it would eliminate the need for a final search for compounds. The splitting takes place at the site of the punctuation marks of the ciphers, so insofar as it is possible, the ciphers split into aggregates, bridges, and interrelated bridging elements. The total number of fragments from the 60 documents analyzed amounted to about 150.

In order to find out how the number of terms varies with the number of documents in the file, 60 more patents were analyzed, and the information was added to the test file. Approximately 70 compound fragments not found in the previous patents were extracted.

The sample sizes reported here were too small to draw meaningful conclusions or to predict anything about the future utility of the file for actual searching.

Comparison of WLN and IUPAC Systems

In 1963, Bonnett reviewed the similarities and differences between the Wiswesser Line Notation and the IUPAC Notation.[19] He pointed out that the designers of the systems had different objectives in mind: in acyclic compounds, for example, Dyson gives indexing precedence to the carbon skeleton, and Wiswesser gives it to functional groups.

While the principles underlying both notations lead organic compounds to be divided broadly into acyclic and cyclic structures, the notations diverge widely in the handling of benzene and acyclic structures. Benzene compounds are grouped with other cyclic structures in the IUPAC system, while Wiswesser treats the benzene ring in a manner analogous to aliphatic structural fragments.

The order of precedence used in the IUPAC system for acyclic compounds assigns to the most prominent position, i.e., the initial, indexing position, the single letter "C" followed by the number of carbons in the senior component, the side chains, any unsaturation, and the substituent functional groups, in an arbitrary hierarchical sequence. This procedure results in a notation structurally analo-

gous to the familiar molecular formula that, for indexing purposes, must perform like a molecular formula.

Wiswesser cites the structural fragments in an end-to-end manner, choosing as the starting point that fragment whose symbol is in latest alphanumeric position. Wiswesser notations tend to mirror the structure as drawn; they, too have some hierarchical characteristics.

The difference in the performance of the two notations on different compounds is striking. In some instances the IUPAC notation will group compounds of similar structure together, whereas the Wiswesser notation results in different notations that do not reveal the identity of the carbon skeleton and that would be scattered in an alphabetized list of notations. Conversely, in some cases the IUPAC cipher effectively conceals the close relationship of particular compounds, whereas the Wiswesser notation for the same compounds tends to preserve their relationships and to group them in alphabetized lists. Of course, any set of rules must subordinate some functional groups, with the result that an index created by a simple alphabetization of the notation will perform poorly when used manually if the subordinated groups are the subjects of searching interest.

In denoting cyclic structures (except benzene), both IUPAC and Wiswesser consider the ring system as the senior component, using a symbol to signal a ring system and an assembly notation for denoting multiple cyclic structures. Both systems define hetero segments after defining the rings and mode of attachment. The two systems then differ in the evaluation of the characteristic to be used in the first and major indexing position of the notation. In the IUPAC system this position is used to classify a ring system broadly as "aromatic" or "saturated"; Wiswesser uses the initial indexing position of the notation to classify the ring system as carbocyclic or heterocyclic. Ring substituents are cited by the IUPAC system in hierarchical order, whereas Wiswesser cites them in locant order.

The IUPAC notation would classify organic compounds into only four groups, namely, "acyclic," designated by "C"; "saturated" ring systems, designated by "A"; "aromatic," designated by "B" (which includes benzene); and macrocycles, i.e., rings of ring systems, designated by "M." Wiswesser notations for acyclic and benzene compounds are divided into many groups according to the initiating functional group symbol; other ring systems are classified as carbocyclic or heterocyclic. Since the number of known organic compounds is in the millions, classification of such a number of compounds into a few main groups in a single index system would result in groups of tremendous size.

The IUPAC notation utilizes uppercase and lowercase letters, normal, superscript, and subscript numerals, and several other

symbols. In 1963, the symbols used for the Wiswesser notations were the 26 letters, the 10 numerals, the ampersand, the hyphen, and blank space. Since then, the slashbar has been added.

Of greater importance than the normal complement of symbols available on the typewriter is the equipment available on accounting and computing machines. Equipment of this nature is required for mechanically manipulating and organizing chemical structures in their notation forms. Generally speaking, standard unmodified accounting equipment offers a limited number of symbols, including uppercase letters, the ten numerals, and perhaps as many as a dozen additional punctuation and accounting symbols.

The limited group of symbols used in the Wiswesser notation is not an unmixed blessing, since it is necessary to assign multiple meanings to some symbols and to depend upon context for distinction as to meaning. This problem is analogous to that of homonyms in natural language.

Both Dyson and Wiswesser feel that their respective notations are more readily understandable than the other. However, it appears likely that familiarity and experience are the dominant factors.

Conciseness is an important attribute of a notation. The IUPAC notation has been shortened considerably by the omission of some punctuation marks. The use of blank spaces in the Wiswesser notation makes the notation somewhat longer but tends to break the notations into "words," which are easier for the human eye to read.

The fact that neither system can be considered perfect for all uses may be reflected in the development of other specialized notations designed for particular problems in information handling.

Hayward Notation

Project HAYSTAQ, a comprehensive computer system for searching chemical information, had its origin in the search for techniques to relieve the burden on U.S. Patent Office examiners.[82] The system was examined in some detail in Survey of Chemical Notation Systems, and its development has been more fully described in a National Bureau of Standards report by Marden.[121] Part of the cooperative research effort of the Patent Office and the National Bureau of Standards (NBS) was spent on the development of a unique and unambiguous linear notation system for chemical structure.[81] The notation system, essentially that of Hayward, was also described in Survey of Chemical Notation Systems. Routines were developed to find either individual complete compounds or simple but common types of chemical structures. The technique relied on matching a

symbol string characteristic of the substructure sought (or an entire notation) with strings within the complete structure notations, provision being made for interruption of the symbol string by arbitrary substrings and for a special type of permutation associated with cyclic substructures.[122] Rules have also been devised for certain types of Markush structures and for coordination compounds (described in a later section of this review).

The line-formula notation system for Markush structures was described by Sneed and co-workers as a supplement to the existing Hayward system developed for specific structures.[169] Markush structures are generic expressions of structures or structure classes; the proposed system is limited to determinate structures of several isolated Markush forms. Investigation has isolated five distinct forms of variable substructure groups. A single Markush structure can have many enumeration patterns because of the flexibility possible in selecting a starting atom. Once a starting atom is chosen, the enumeration and ciphering are similar to the general rules of the Hayward system, except that special rules and symbols are needed for the Markush nodes and the frequency variables.

The NBS staff has reported on an extension of the HAYSTAQ system to search for embedded substructures in a retrieval system based on chemical structures.[108] The predominant technique is one of tracing paths from selected points in both the structure input for checking and the structure chosen for comparison on the basis of prior screening, and comparing these paths at each point along the way. The search routine, TOPKAT, is limited to consideration of topological properties of structure only. The current format is adapted for use on the UNIVAC 1108 computer.

Additional developments in the Hayward notation are discussed later in this report.

Hercules Incorporated Notation

A linear notation system has been installed at Hercules Incorporated to handle chemical structures of interest as pesticides.[164] The various codes possible, depending on the viewpoint of the indexer, are stored and sorted on punched cards.

An index of chemicals of interest as pesticides must be designed to express the subject content of chemical structures from the following viewpoints: (1) the whole structure, or essentially the whole structure, (2) the toxophoric group or groups, (3) other functional groups or moieties, (4) relationships between parts of the structures within a toxophoric class, and (5) relationships between toxophoric classes.

The notation system developed at Hercules for the indexing of pesticides has features of the Dyson, Silk, and Wiswesser systems, plus several new features. The cipher is started from any functional group or moiety in the chemical structure. Positions of substituents on rings are not designated. In an extremely large file, of course, positions of these substituents would need to be designated. Within a given toxophoric class of pesticides, however, position designation is an unnecessary refinement. Ciphers are written in the order of preferred toxophoric group, followed by that portion of the molecule containing other toxophoric groups or moieties of interest.

The same order of citation is followed when a second index entry is made for the same compound, It it is important that a compound be considered from more than one viewpoint, the necessary number of entries may be made in the index. The degree of detail in the cipher and the depth of indexing are influenced by the number of chemicals to be indexed, the complexity of the chemicals, the interrelationships that have to be correlated, and, most important, the needs of the users. The notation code, by design, is as flexible as any machine card code in denoting functional groups and moieties, with the additional advantages of showing relationships between functional groups and moieties and of being particularly suitable for the production of machine printouts.

LINCO Notation

The LINCO notation, devised at Shell Internationale in The Hague, the Netherlands, is described quite clearly in Survey of European Non-Conventional Chemical Notation Systems, where it was assessed as "one of the more promising notation systems observed in Europe." At the time of that evaluation, there had been only a limited amount of computer searching with the notation. Since then, a paper presented at the national meeting of the American Chemical Society in April 1964, and later published,[22] gives more details of the system's development.

The system is based on treating structural formulas as networks containing branching points (points where three or more series of bonded atoms other than hydrogen interlock), giving to these points an arbitrary number, and writing down the series of atoms between two points of that type ("bridges") together with the series of atoms extending from such branching points but not meeting others ("chains").

Although the data in any of the storage lists that can be obtained are sufficient for searching purposes, more rapid searching without

computation steps would be possible if the list could be standardized by revision of the reference numbers alloted arbitrarily to the branching points. For this purpose, each of the branching points is given a special number, known as a yardstick number, calculated for each branching point in each step. The counts, in short, are: all incoming connections, all outgoing connections, and all connections going to the next level. The results are put in weighted order and used as a yardstick number for the top branching point of the network investigated, to be compared with the similarly obtained yardstick numbers for the tops of the other networks.

The simplicity of the system for the user is obtained at the cost of relatively complex, albeit straightforward, machine actions, since the machine has to substitute something for all the details (such as priority rules) which a human using another system ordinarily has to take care of.

Substructure searches can be made in several ways, including the special case of the search for rings. Those present can be deduced from the matrix used: it is not necessary to ascertain the smallest rings present. If desired, every ring present in a compound can be filed, together with the storage list, as a series of bonded atoms. If for complicated structures this would mean too many rings, those noted can be restricted to medium-size rings only. By adding the reference numbers encountered in the ring, the place of substituents is easily ascertainable.

An extended matrix forming an atom-by-atom connection matrix can be used to translate the information stored by the LINCO system into other codes (such as the IUPAC notation). It may even be possible to base a system of unique names for compounds on the standard representation developed.

Polymeric Species Code

Lissant described a method of defining polymeric species and of differentiating between species and between individual members of a species.[117] The technique consists of mapping the properties of members of a class of polymers in a suitable composition space chosen in such a way that the interrelationships are readily displayed. Monomer and product proportions, physical properties, and reaction conditions are considered. With such a system, each polymer has a unique "name" in a sequential notation, and all data pertaining to that polymer can be filed under that index entry. The system is amenable to searching for specific polymers, subclasses of polymers, or polymers with specific properties.

D. TABULAR AND GRAPHIC REPRESENTATIONS

The preceding sections of this report have reviewed the methods of recording chemical structural formulas that involve fragmentation into atoms or groupings of atoms and generation of two-dimensional strings of characters. This section is devoted to other means of complete representation of chemical structures, for example, atom-by-atom and bond-by-bond listings of the parts of a molecule. For some aspects of computer processing and output, the atomic groups and connectivity information must be separately available and yet form a unity, known as a connectivity table. Various possibilities exist for composing and storing a connectivity table.

Developments in the field can be reviewed with reference to an outline proposed by Opler of eight different schemes that have been used with computers[144]:

1. Diagrammatic representation A matrix of symbols when written out row-by-row on a line printer produces a crude diagrammatic representation. Special programs have been written to recognize structures (and substructures) by "geometric matching." The Survey of Chemical Notation Systems described the Monsanto system of Waldo and DeBacker, which is an example of this approach.

Jacobus and Feldman at Walter Reed Army Medical Center (Washington, D.C.) have developed a chemical typewriter by means of which the identity of a struck key is recorded on paper tape, along with the coordinates at which the impression is made on a graph or grid.[130] The typist may type the structure in any sequence. The initial paper-tape record is fed through a small computer that acts as a sorting machine to sort out the strokes, arrange them in sequence, and punch out a new paper-tape record. This new paper-tape record may then be fed back to the typewriter, which will then type out the structure line by line. (The Walter Reed System is described in more detail later in this report.)

2. Use of dummy connectors The arbitrary numbering of nodes allows a computer to recognize the connective properties of a sequential list of atoms (nodes) to which these artificial (dummy) numbers refer. Again, the Survey illustrated the method in descriptions of the systems of Mooers and Norton and Opler.

3. Juxtaposition and concatenation It is possible to store and manipulate ciphers (such as Wiswesser's) in computers. Recognition rules may be programmed. Examples are described in an earlier section of this report.

4. Node-pair lists Graphs have been represented as ordered lists of node pairs, with Newell-Simon-Shaw list structures.

5. Lists of symbols Opler represented structures as linear, branched, and cyclic Newell-Simon-Shaw lists in which the nodes contain symbols for the element, while the list structure (as represented by computer elements) represents the structure of the graph.

6. Tabular representations Structures can be represented as a table in which the entries are numbers representing atoms along with the entry numbers of atoms connected with them, as described, for example, by Dyson and others.[50] This method is akin to Number 2 above.

7. Incidence matrix representations In the representation as a matrix of 1's and 0's, the so-called incidence matrix is square and of the nth order, where n is the number of nodal elements. The 1's represent connections between nodes; the 0's represent no connection. Sussenguth treated the matrices from a chemical viewpoint.[178]

8. Polish Notation In "punctuation-free" notation as developed by Hiz[83] and Eisman[55] after the "operator-free" algebraic notation, simple algorithms give the encoding rules. The decoding rules are more complex and preferably left to a computer.

"Notch and punched card equipment made the fragmentation technique practical. The computer . . . offers the capability of processing the type of detailed record produced by fragmenting all the way to the atom and interatomic bonds while retaining the context between these fragments. . . ."[20] Few and relatively simple rules are required for human manipulation at input. Functional groups and other structural features need not be identified as such. The coder is not concerned with any priority, ranking, or sequence problems. As is generally the case with systems, chemical knowledge is required for the initial preparation of the structural formula. However, people with less skill can complete the input steps to the machine. The machine can be programmed to do considerable checking and editing and can analyze the structure to produce a canonical form. "It is only since the introduction of machine methods . . . that there has been any prospect of exploiting the topological method. . . ."[130]

The principal motivation for using these representations is the possibility of utilizing computers to operate upon them. According to Opler,[144] some of the more significant operations include:

1. Determining if two graphs are topologically identical (isomorphic). This problem is trivial in theory but difficult in practice. For large structures, the time required to prepare all permutations and to compare them with the original is prohibitive. The problem of minimizing the effort of determining isomorphism has been attacked by using manipulations of incidence matrices and by matching node pairs heuristically. Such work is discussed later in this section.

2. Determining if one graph is included within another. This problem is the one most commonly attacked because it involves the manipulations required to determine if a chemical substructure appears in a larger structure.

3. Transforming one representation of a structure into another. Transformations are an intriguing possibility. Matrices or connectivity tables as atom-by-bond-by-atom representations of structure might be types of common or basic representations to which other types could be converted. Thus, two different notations or two different connectivity tables could be interconverted, a subject discussed in more detail later in this report.

4. Determining if a structure is well formed. A graph is well formed if each node uses the exact permitted number of branches and all nodes in a single structure are connected. In addition to these mathematical requirements, there may be requirements stemming from steric effects, valency, and other chemical factors. The incidence matrix representation lends itself well to formal procedures for checking whether a structure is well formed.

5. Generating well-formed topological representations. This task should be relatively easy for a computer, given the rules for well-formed representations.

6. Converting from diagram to topological representation. With the development of mechanical aids to feed diagram representations into computers, considerable attention has been focused on algorithms for converting such geometric representations to topological representations.

Du Pont System

A Chemical Structure Storage and Search System (CS^4) based on topology has been installed for the use of four centralized information groups within Du Pont.[84] Computers are employed in both updating and searching the files.

Index terms selected for each document are stored in one of two inverted files. The General Term File uses alphanumeric fields to identify nonchemical terms with associated document accession numbers. The Compound File uses numeric compound numbers to identify terms. Both files provide for link and role indicators. Supporting these files are several other computer-maintained second-level index files. The Thesaurus File stores each valid term in the General Term File. The Fragment File stores chemical compound descriptions in predetermined functional groups and auxiliary descriptors. The Registry Files store structures of chemical compounds.

Input to the General Term and Compound Files is in the form of records containing an index term, roles, accession number, and link. Connection-table input is submitted outside the input stream of the Document System, resulting in the development of the linkage approach rather than direct integration of programs.

The CS[4] system stores the topology of chemical structures in the form of connection tables. A registry number is assigned to each input structure; registry numbers of compounds containing specified substructures are retrieved.

A chemical structure must be known to be registered. Input to this Registry System is in the form of connection tables describing the structure of a compound in terms of atom–bond connections, element symbols, and bond-type codes. Chemical structures are keyboarded with a Mohawk 1181A onto magnetic tape, converted by program to connection tables, and sorted. Input structures are then checked for errors. Error messages and the input connection tables are printed. Bonds in rings are detected and marked with a special set of codes.

A unique (canonical) form (numbering) is computed for each input compound. The structure is renumbered according to an algorithm developed by Morgan[133] that ensures computation of the same numbering regardless of the order in which the atoms were input. The input connection table is reformatted to a compact form that stores atom connections, bond codes, and element symbols in separate lists. A screen file containing information extracted from the connection-table record for each compound is updated as the last step in each updating of the CS[4] Registry System update.[84]

Inquiries for the Substructure Search System include a substructure drawing, identification information, screens, and substructure coding. The substructure is composed of one or more groups to which a logical operator—"AND," "OR," or "NOT"—is assigned. The combination of groups defines the substructure inquiry. Screens prepared by the inquirer specify minimum features a registry compound must contain to justify attempting iteratively to map the substructure onto the compounds. Screens permit specification of atom and ring counts, bond types and counts, molecular formulas, and element–bond–element triplets.

Inquiry forms are keyboarded directly onto magnetic tape. An edit program reprints each question and diagnostic or error messages. The Screenout program compares the screens prepared by the inquirer with information stored on the screen file. The Substructure Search program consists of two main phases. In the first, each substructure inquiry is compiled into a series of machine-language instructions that will actually perform the search. Phase Two performs the substructure search. The compact connection table for

compounds that passed the screens for one or more questions is expanded to a one-atom-per-row connection table. After each compound is read and expanded, control is transferred to the compiled questions. If the iterative search is successful in mapping the inquiry onto the compound structure, the registry number and inquiry number are written on an output tape. An editing routine either prints registry number answers or formats the answers for the Document System search.

With a few minor exceptions, all programs in the CS[4] Registry and Substructure Search Systems were run on an IBM 7010 computer under the PR-155 Operating System.[84] The IBM 7010 assembly language, AUTOCODER, is used for all major processing programs. All output edits and the substructure search input edit employ COBOL. The registry files at the time contained approximately 55,000 nonpolymers and 14,000 polymers. The collection was growing at the rate of 18,000 annually. Total input cost per compound was reported to be $0.882.* Search computer costs are a function of the number of questions run in a batch. Searches of the nonpolymer registry file cost $54 per search.* Searches of the polymer registry file cost somewhat less. The cost figures do not include technical time to prepare structures and frame searches or clerical efforts, such as filing of molecular formula cards and connection-table sheets.

Major modifications and extensions of the Chemical Structure System have been incorporated to permit registration and substructure searching of coordination compounds, complexes, and polymers. Two quantities, called connection number and oxidation state, are defined for each atom. Connection number is the sum of bond values, including hydrogen, attached to an atom. Oxidation state is an attribute that describes the number of valence electrons donated to or accepted from a ligand. In a substructure search, specification of unusual connection numbers or oxidation states or unusual configurations permit retrieval or exclusion of corresponding atoms in coordination compounds. The inquirer may also specify that only standard values are acceptable.

Linear polymers are structured for CS[4] by coded representations of one or more significant repeating units (SRU's) and end groups. As explained in an earlier paper,[71] the problem with polymers comes from the chemist's inability, in many cases, to describe accurately the structural representation of a specific polymer chain, and the undesirability of requiring the system to handle something as long and redundant as a polymer chain, even if the chemist could. The solution is to input each monomer in the form in which it would

*Cost estimates should be viewed with caution, as noted in the Preface.

exist in the polymer chain linked to a dummy central atom.[84] End groups, if any are known to exist, can also be linked to this dummy atom. Additional conventions, not yet implemented, have been developed for grafted, cross-linked, and postreacted polymers.

Polymer structures are drawn on the CS^4 input forms. Connection-table records are processed by the computer programs used to process nonpolymers. After computation of the unique form, polymer structures are added to a separate master-file tape. Relatively few changes were required in the program that computes the unique form for each input compound. To compute a unique numbering without regard to bond descriptors, two copies of the bond list must be maintained. The second version, which has single bonds substituted for descripted bonds, is used in the computations. In all other respects, polymer structures are registered in the same manner as nonpolymers.

The polymer version of the CS^4 Substructure Search System maps a sequence of atoms and bonds onto a structure, even if the complete sequence is not found within a single significant repeating unit. End-group searches are coded identically with searches of SRU's. The inquirer must indicate that a substructure search is intended for mapping only against end-group atoms.

Implementation of CS^4 at Du Pont required a series of compromises to permit use of the available programs. It was desired to make the system operational as soon as possible and to gain experience with use of a topological chemical structure system.

ChemSEARCH

The Colgate-Palmolive Research Center has an operating topological system covering structures of company interest.[73] Since the input to the system is by clerical coding, the chemist inputting a structure to the file is required to write all its atoms and all bonds between them, but with omission of hydrogen bound to noncarbonyl carbon. The resulting structure is just as intelligible to the chemist as the more customary condensed forms.

The conventions were established, partly to require some thought on the part of the chemist entering a structure, to emphasize the generic relationships of families (for example, salts and their parents), and to establish some consistency of input format. Other goals were simplicity and clarity of presentation, ease of coding, and avoidance of errors. However, given any correct depiction of atoms and bonds in a topological network, all other conventions can be ignored, and the structure is retrievable from the file. It is claimed

that the flexibility of the ChemSEARCH system allows the professionally trained chemist, with sufficient thought and knowledge of possible alternative representations, to formulate a single question that will retrieve not only identical (perhaps mesomeric) structures entered with varying but accepted bonding, but even closely related structures, such as tautomeric keto-enol pairs. Since this system is designed primarily for use by chemists interested in structure-activity relationships, certain conventions for salts were adopted to increase the ease of obtaining an entire related family. Of greatest importance, however, was elaboration of the concept of generalized atoms and bonds, which permits generic questioning of the file with a high degree of sophistication.

After a compound is drawn by the chemist, with all bonds explicitly indicated and appropriate hydrogens omitted, the structure with atoms numbered is entered in writing on a coding form (average time, 3 minutes). The data, including number, name, empirical formula, and structural formula, is punch-typed on paper tape. The Syntax Checker program for error determination (on paper tape) is inserted into the computer, followed by the paper tape corresponding to the typescript. The Syntax Checker makes 47 different internal checks on the correct format and content of the coding, such as back coding, atoms bound to a subject atom that are not also subjects, correct empirical formula for a given structure, equality of bond between two atoms, and even whether the atomic codes used have been approved by IUPAC. About 4 percent of structure codes are found incorrect, and another 4 percent of more easily correctable typographical, format, and counting errors are detected. Once such errors have been corrected, the typescript set of codes can be used to reconstruct the molecule, which is then checked visually, preferably by the chemist, for identity with the original structure submitted. In practice, however, only about 0.1 percent of the structures have been found to be wrong at this point. Therefore, structures are usually inserted into the main file immediately after syntax checking and deleted and replaced later if necessary. Another method of achieving a final check is through the regeneration of conventional structures by the Walter Reed Army Chemical Typewriter (described earlier in this section). To this end, a cooperative project has succeeded in transforming the codes of the Army Chemical Typewriter into the complete topological description of a structure searchable by the ChemSEARCH system. This interface also allows commercially available coding typewriters to be used for visual input through modification of the HECSAGON system of Horowitz and Crane.[86]

When one or several sets of compounds are ready for insertion into the main file, the File Maintenance program is used. The file

is then ready for searching by the Question Reader program that compares structures; any reasonable number of questions may be asked simultaneously. If the computer cannot handle them at once, it stores some of them for a second run. Format for most questions is identical with coding for file input, and any subsection of the total code except the name may be used for searching. Inclusion of a greater fraction of the total code record leads to greater and greater specificity in a question. Useful searches can be made on the empirical formula, particularly in the case of inorganic compounds, which have generally been entered without structure. Since there are limited possible structural combinations of inorganics, a fairly specific answer can be obtained even without structure.

To achieve the structure-activity correlations desired by the synthetic chemist, a computer file, keyed by number to the structure file, is being constructed to list alternative names (including trade names), sources, literature references, and qualitative and quantitative properties. A search for a given value of a certain property (or values exceeding the given value) would then elicit a set of accession numbers. The corresponding structures are readily obtained from the structure file of the ChemSEARCH program. Conversely, a related set of structures obtained by ChemSEARCH could be used to obtain their properties for comparison.

Penny Connectivity Code

A proposed computer method for handling the problem of recognizing chemical structures is based on set theory whereby sets are generated from graph-theoretic and chemical characteristics of the structure.[150] A chemical structure can be regarded as a graph consisting of points (nodes) and lines (edges) connecting these points. A node represents an individual atom in the structure and its "node value" is that specific chemical element associated with the node. The edges indicate the connectivity or bonding between the atoms. In the case of double or triple bonding, the degree of bonding between two atoms can be represented either as an attribute of a single edge (an assigned multiplicity) or by the presence of multiple edges. However, the degree of bonding covers but a small set of possible values. The same is true of the set designating chemical identity because of the predominance of four elements (C, H, N, and O) in organic compounds. It would therefore appear that the development of a convenient means (notation) for describing graphical relationships between atoms will provide an information set with a high level of differentiating power.

The connectivity code (through, for example, three levels of connectivity) is made up of three parts: groups, the number of characters within a group, and the numerical value of these characters. Each branch from the subject atom to the first connection level corresponds to a group. The number of branches to the second level from each first-level node is designated by the number of characters within each respective group. The number of branches from each second-level node to the third level is indicated by the numerical value of the characters described above. The generated connectivity would initially be represented as an arbitrary arrangement. To express the code in a canonical form, both for uniqueness and to facilitate comparisons of identity, the characters within groups may be arranged in descending numerical order, and the groups themselves, in descending order by the most significant character.

Importantly in a graphical sense (and in practical application), the connectivity code handles the problem of recognition of inverted or rotated configurations. A structure and its mirror image will have identical codes. The relative position of the groups does not necessarily reflect a graphical configuration relative to any preconceived coordinate or geometric system; rather, the code is ordered on an arithmetic basis.

An application of the Penny connectivity codes was suggested for the Chemical Information and Data Systems (CIDS), to be used in conjunction with BATCH numbers (explained earlier in this report) and molecular formulas.[208] The proposed organization of the files calls for a single control word consisting of BATCH Number, Molecular Formula, and Bond Summary, so that a detailed examination of a given compound is required during whole compound searching only when it has the same control word as the query compound. When this is the case, a more detailed examination can be made utilizing the Penny connectivity codes. A different approach would be required for fragment searching. The fragment control word must be in all cases either identical with or imbedded in the master-file control word before a detailed examination is required.

Topological Ciphers

Ballard and Neeland have pointed out that the rules for topological encipherment are generally more complex than those for fragment codes.[8] Their effort has been slanted toward separating the intellectual tasks from the manipulative tasks by placing the latter burden upon a computer. The objective of their technique is to provide an automated means of retrieving unambiguously from among a large

collection of organic and inorganic chemical ciphers those possessing in common one or more desired structural similarities, while at the same time conserving human effort on both the data input and the preparation of search specifications.

The approach of topological description for enciphering the compounds results in a longer initial cipher than most other coding schemes, but it furnishes the computer with complete operating data and enough redundant statements to permit automatic checking of the self-consistency of each cipher. A second important part of the input processing is the preparation of various data of summary type for coarse screening, such as the empirical formula, the number and types of ring structures, if any, the number and types of acyclic groups, and the number of each type of bond present. Failure of a chemical cipher to meet any of the coarse screens for a given question will result in bypassing the detailed search of that particular cipher for that inquiry. A third part of the input processing is concerned with a reorganization of the input cipher to decrease the number of steps necessary in the computer program to keep track of the atom-by-atom search that would follow successful passing of the coarse screens.

The desired gain in input accuracy through reduction of the level of intellectual effort is paid for in terms of a rather complex input and error-detection program. There are, however, compensating benefits to be derived from this program. The summary data for the coarse screens are simply tallied as they are recognized by the input routine, with very few additional program requirements. With a relatively small further sophistication of the input program, the input data can be converted from the restricted viewpoint of the input cipher, which treats each atom as an encoding point, to the next higher level of organization of the facts. Each interrogation will be prefaced with the minimum acceptable atom count and the minimum acceptable number and types of rings, chains, and bonds. These numbers are checked against the empirical formula and the summary data for screening prior to an attempted detailed search.

No reports of operating experience with this system have been found.

Tree Structures

Eisman proposed another method for describing chemical structure graphs, in particular tree structures. "A tree is a graph in which there is no cycle."[55] Several schemes exist for describing trees in a linear fashion, rather than in the customary two-dimensional man-

ner. For example, everything within a pair of parentheses is considered bonded to the element just in front of the left-hand parenthesis symbol defining that parenthetical level. Since the character representing the type of atom must also contain information regarding the valence, several different symbols must be used for an atom that can assume different valences. A computer subroutine can be written that will determine how many bonds are connected to a particular atom and select the preassigned codes for the form of the atom having that particular valence. Similarly, the practice of dropping "H's" should present no problem to a mechanical system. In this case, the atom to which the "H's" would have been connected will have a different code.

Using such conventions, the parentheses can be eliminated from the linear representation, since the information they contain is implicit in the order of the nodes and the table of degrees. This form is called parenthesis-free (Polish-type) notation.

Hiz proposed rules for obtaining a parenthesis-free notation recorded as a two-dimensional graph.[83] A unique representation of the compound is not obtained since the notation depends both on the orientation of the structure as originally drawn and on the choice of starting point.

Welch demonstrated the potential of Polish-type notation as an internal symbolic language for substructure search.[204] The initial compound string is manipulated into a form in which the substructure string appears as a contiguous portion. The compactness of the Polish notation allows backtracking by generating all the compound string variations matching the substring up to the examined symbol. Conversion to the Polish form from other existing representations probably could be done by computer without requiring a canonical form of the derived representation. The Polish output of substructure searches has the advantage of showing in what ways the substructure occurs in the compounds retrieved.

Sussenguth also proposed the use of tree structures for processing files.[177] In many data-processing problems, files must be not only searched but altered.

Binary search techniques are efficient for searching large files, but the associated file organization is not readily adapted to the file alterations. Conversely, a chained file organization permits easy alteration but cannot be searched efficiently. A file organized into a tree-like structure may be both searched and altered in times proportional to $s \log_s N$, where N is the number of file items and s is a parameter of the tree. Optimizing the value of s leads to a search time which is only 25 percent longer than for the binary search. The tree organization employs two data chains and may be considered to

be a compromise between the organizations for the binary search and the chained file.

Barnes suggested that although trees are both important and useful, one must be prepared to deal with graphs that are not trees because certain phenomena with cyclic features cannot be adequately represented by a simple tree.[15] He suggests a graph-theoretic language as one approach to the direct representation of natural language.

HECSAGON System

A search routine has been outlined in principle (though not fully programmed) that permits bond symbols to be either horizontal, vertical, or diagonal and, with some concessions to expediency, may preserve some of the conventions of classical structure notation. It is called the Horowitz-Eastman-Crane Symbol-Array Governed-by-Orthodox-Notation, or HECSAGON, System.[86]

For reconstitution in the matrix of computer "memory," symbols must be arranged in rows and columns and must stay within the gamut of the card-punch keyboard. All capital letters and on-line digits are available for conventional use. Lowercase letters, superscripts, and subscripts could be contrived by a two-character code and a document-writer or slow printer, but they are not really necessary. Any digits could be printed as subscripts by a hypothetical full-speed "chemical" printer, which is not, however, required for utilization of HECSAGON.

Eleven special characters are available. Not all are the same on different keypunches, but as long as the punch patterns always have the same chemical significance, anyone can use anyone else's punched cards. For the sake of clarity, six special characters can be assigned to describing the spatial orientation of single bonds. Orientation of double and triple bonds can be deduced from the location of the atom symbols connected. Electron deficiency, the lone electron, and the skeleton vertex also deserve special characters. Electron excess may be symbolized as an unshared single bond.

Like the chemist, the HECSAGON search routine seeks only topological relationships in structural formulas. A notation may therefore be oriented and distorted at will. There is no need to anticipate the style, orientation, or distortion of search requests.

Formulas are to be written on paper ruled in squares, one character per square. A one-letter atom symbol is surrounded by eight squares, a two-letter atom symbol, by ten squares. In general, these surrounding squares are to be used only for other symbols pertaining

to the central symbol—a number multiplying it or showing its weight, or the charge on it, or bonds attached to it.

Applications of the HECSAGON system, in the fragment-code system at Abbott Laboratories and the ChemSEARCH system at Colgate-Palmolive Research Center, were described in preceding sections of this review.

BRAID System

It has been suggested that where the descriptors for a search may be any or all of the components comprising the entries, and where it is also necessary to retrieve the entire entry rather than merely a reference to it such as a document number, one might take the concept of an inverted term file, expand it into a total concordance, and then code the original text by reference to these terms. It is claimed that if the terms themselves are treated as documents composed of letters for which another inverted file is formed, and if a number of intermediate reference levels are similarly generated, the resultant BRAID (Bidirectional Reference Array Internally Derived) displays surprising advantages in storage requirements and in retrieval time over other schemes.[157] Datatrol Corp. suggested that BRAID could handle reference files of 100,000 to 1,000,000 chemical structures with complete flexibility in dealing with "new" combinations of atoms, could automatically and economically update the reference file for any additional structures, and could locate and output with extreme speed all entries corresponding in whole or in part with any question while excluding all extraneous entries.[158]

The starting point is the topological representation of the chemical structure. The retrieval problem is to locate any occurrence of an arbitrary portion of such a structure. Two methods of searching a file are proposed. One involves examining each entry for a correspondence with the question. The second involves listing questions or their components in advance, with corresponding file entries listed after each. Where the question can be any subset of a chemical structure, however, one faces a choice of either forming an excessive number of possible terms or of leaving an undue amount of work for the tedious direct search. BRAID may be viewed as an attempt to solve this dilemma by having the computer derive a bidirectional tree in which both methods of search can be performed concurrently with immense savings of time. Briefly, the array consists of a family of "entries," each entry consisting of three sectors, which might be called incidental, definition, and reference sectors. The incidental sector I contains any information about the corresponding term other

than the chemical structure itself. The definition sector D must constitute a sufficient description of the entry-associated structure as a properly connected set of substructures, together with the remaining bonds; and the reference sector R lists those entries in which this one is directly contained.

Each entry is a node for two overlapping symbol trees that run in opposite directions. The conception of BRAID does not require either of these trees to be completely general, however; each node of the definition tree, or D-sector, is of fixed length known at the formation of the entry, and therefore stored in sequential cells. The R-sectors are subject to constant growth, but can be more conveniently chained by blocks than by individual items. The first operation of the computer system involves reading some representation of the chemical structure. Each symbol will relate to a bond or to an atom; the first step is to enumerate correctly all the bonds and to associate the remaining symbols with the proper set of atoms. The input structure finally appears in the same form as the rest of the BRAID. The search procedure will consist of a systematic progression up the reference tree of the BRAID to locate or construct an entry identical with the input structure (during update) or to locate every such entry (during retrieval). A large number of paths may be eliminated at each step, an obvious advantage with large files.

As the search progresses from level to level, the complexity of the structures becomes greater. The search continues until either a structure identical to the input structure is found or no paths exist for continuation of the search. If the search is terminated without a match being found, the input structure is considered to be a new entry for the file, and the procedure for updating is carried out. When an exact match is found, it is tagged for output. If the match occurs at an entry that is not a complete structure (that is, if the input structure was only a fragment), all larger structures containing the matched entry are selected for output.

No reports of operating experience with this system have been found.

DENDRAL System

Lederberg proposed a system for computer construction, enumeration, and notation of organic molecules as tree structures and cyclic graphs, called DENDRAL (from DENDRitic ALgorithm).[106] The principle distinction of DENDRAL is its algorithmic character. Each structure has an ordered place, regardless of its notation.

DENDRAL aims (1) to establish a unique (i.e. canonical) descrip-

tion of a given structure, (2) to arrive at the canonical form through mechanistic rules that minimize repetitive searches and geometrical intuition, and (3) to facilitate the ordering of the isomers at any point in the scan, thus also the enumeration of all of them. The program has two aspects: a notational algorithm, transposing a stated structure to its canonical form, and a generative algorithm, successively building each of the hypothetical structural isomers for a given composition. The one standardizes the representation of a given structure to confer its unique location in the dictionary, while the other generates a dictionary of all possible structures.

A computer program written to implement DENDRAL can generate all the structural isomers of a chemical composition.[179] The generated structures are inspected for forbidden substructures in order to eliminate structures that are chemically impossible. The program incorporates a memory so that past experiences are utilized in later work. A list processing language, LISP, was used to write the program, which is made up of over a hundred separate LISP functions calling each other to perform certain tasks relevant to different parts of the structure generation. Most of the functions are simple utility programs. The main logic is contained in eight or ten primary functions.

The DENDRAL program is constantly being modified as new and better procedures are conceived. The basic structure-generating functions are written independently in such a way that new supervisory functions can be easily inserted onto the system. It is hoped that the DENDRAL program will eventually be able to benefit from the user-on-line characteristics of a time-sharing system in order to extract knowledge from chemists and other users of the program.

The staff of the Stanford Artificial Intelligence Project reported on another computer program, HEURISTIC DENDRAL, that can formulate hypothetical molecular structures from data consisting of the mass spectrum and empirical formula of an organic chemical compound.[29] The foundation for the program is Lederberg's DENDRAL algorithm. The program is heuristic in that it discovers the plausible hypotheses according to rules learned from chemists constituting a small subset of the total set of hypotheses. This program is also written in LISP for the PDP-6 computer at the Stanford University Artificial Intelligence Laboratory.

Atom Connectivity Matrix

Spialter proposed a system based on the atom connectivity matrix in which each distinct chemical structure is represented as a unique

mathematical expression in the form of a characteristic polynomial.[172] A representation of molecular structure based only on atoms and interconnecting bonds is probably the simplest mathematical transformation that can be devised. A clear-cut one-to-one correspondence with the chemical structure is obtained with a minimal number of rules, utilizing essentially no additional symbols beyond the usual chemical ones. The Atom Connectivity Matrix (ACM) belongs to this class of representations.[179] It is a matrix composed of elements $\underline{a}_{\underline{i}\underline{j}}$, where $\underline{i}$ and $\underline{j}$ are integers defining the row and column, respectively, at the intersection of which $\underline{a}_{\underline{i}\underline{j}}$ is located. Along the diagonal, where $\underline{i} = \underline{j}$, are placed symbols for the atoms, radicals, electrons, charges, or constituent groups making up the molecular formula. Order of assignment is unimportant. In the respective off-diagonal position, $\underline{a}_{\underline{i}\underline{j}}$, where $\underline{i} \neq \underline{j}$, is placed the "connectivity" (bond order, energy, force constant, directional cosines, or any other bond parameter) between the atom at $a_{\underline{i}\underline{i}}$ and the one at $a_{\underline{j}\underline{j}}$.

The ACM is defined independently of atom-ordering priorities, numbering conventions, or language. For large molecules, it may appear as a formidable array. However, a significant saving may be effected in the size of the ACM by noting that its greatest application will probably be to organic molecules, both because of their overwhelming numbers in the literature and because of their large numbers of isomers. Since hydrogen atoms constitute a large fraction of the atoms in most organic structures, and since their bond-order connectivity has a value of unity, it has been found possible and convenient to omit them from the ACM with no significant loss in generality for a majority of applications. In a later atom-connectivity matrix characteristic polynomial (ACMCP), it was found more useful to develop a geometrical notation by extrapolation and generalization from the original concepts.[174] This development permits the chemist to bypass the ACM and to construct the coefficients of the ACMCP for the homoatomic case directly from the pictograph, retaining such classical geometrical components of structure as rings, multiple bonds, and the like. No reports of operating experience with this system have been found.

Cossum and others have summarized some advantages of the tabular encoding of chemical structures.[41] The codes can be manually generated at a reasonable cost and by relatively untrained personnel from structural diagrams; they completely represent the topology of two-dimensional structural formulas; they are nonhierarchical, thus no aspects of structure are subordinated to others; they can serve as the basis for direct automatic computation of molecular formulas for checking purposes; they can be searched by relatively simple programs; and they can be transformed automatically into unique forms.

Disadvantages are that the codes are not compact (some of the advantage gained at search is offset by the length of the record), and they are unintelligible to chemists (thus the system must include a parallel store of names, structural diagrams, or notation).

Standardization of Graphs and Tabular Codes

In problems of information retrieval by computer, the question of identity and near-identity between enumerated strings of information emerges in different guises time and again. Problems characterized by the need to "permute and test against all possibilities" often yield to algorithms exploiting the characteristic of ordering reflected in the following concept: If the factors that characterize a set can be exhaustively and discretely enumerated, the set of enumerated factors can be ordered invariantly by a routine that preserves the characteristic factors. Sets or strings so treated are often amenable to powerful and rapid solutions of identity and near-identity. According to Wilson, the problem of graph isomorphism can be treated by enumerating the characteristics that identify a given graph and by so ordering them as to lead to one and only one enumerated set.[207] Hiz suggests a linearization of chemical graphs, neutral as to the chemical properties, at a sacrifice in formula length, giving a closer representation of the graph itself.[83] The system of rules he proposes gives for each graph and each path through it a formula that is a record of the path; the rules have a large degree of flexibility and give as many ciphers for a graph as there are paths through the graph.

Unfortunately, graphs do not satisfy all the requirements for representation of molecular structures. In the theory of graphs, two graphs are equivalent if they differ only by the angles between the edges. Chemists are familiar with situations in which two graphs considered equivalent by mathematicians are taken to represent two different structures. A typical example is provided by cis and trans isomers.

There remains the problem of how to discover that two different ciphers are synonymous, that is, that they mark two paths of the same graph. Two ciphers that are synonymous produce short-form matrices differing only in numbering of the nodes; the cross-references will be the same. If a one-to-one mapping of elements of one short-form matrix into the elements of the other can be found, such that the mapping transforms the first short-form matrix into the other, then these matrices and the ciphers that produced them are synonymous. The problem of finding such a mapping requires some

ingenuity in order to avoid long computations and an unmanageable amount of backtracking. An exact matching procedure would not be very helpful because many different ways can be found to express the same ideas or requests. What is needed instead is a procedure permitting inclusion of partly unspecified information and providing for possible relaxation of the various conditions that render a complete match impossible at any given time. Such graph-matching techniques are applicable to the comparison of chemical molecular structures. Structure-matching programs usually operate on a node-by-node or a piece-by-piece basis. In the former approach, the nodes of the two structures are compared one at a time until either the complete structures match or an incompatibility arises; in the piece-by-piece approach, a dictionary of basic substructures is used to break a given structure into pieces that are then matched as a whole. Neither of the techniques works well for any but the simplest structures: the node-by-node method usually requires extensive backtracking, while the piece-by-piece approach suffers from the lack of a standard, well-defined method for breaking a given structure into substructures.

Sussenguth and Salton reported on a topological structure-matching procedure programmed for the IBM 7090 computer that does not depend on a specific ordering of the nodes or on the presence or absence of certain specified substructures.[156] Little or no backtracking is required, and the method can be used to detect complete matches as well as partial ones. The basic idea is to determine certain simple properties of the nodes of the two structures to be matched and to equate those subsets of the nodes exhibiting equivalent properties. A standard procedure is then used to form new matching subsets and to break down already existing subsets into sets with fewer members. The procedure is completely determinate except in cases in which it is necessary to resolve certain symmetries in the connection pattern of the nodes. In such a case, a guess (assignment) is made as to the correct solution—a guess that will later prove to have been right or wrong, and if wrong, may require some backtracking. In most practical problems, however, little backtracking seems to be needed. Computer experiments indicate that this topological procedure is much more efficient than either the node-by-node or the piece-by-piece approach.

The algorithm proposed for making the necessary comparisons consists of two basic parts[178]: (1) the generation of pairs of corresponding subsets of nodes and (2) the partitioning of these subsets to determine correspondence between individual nodes. In a typical example, node values, branch values, and connectivity patterns of nodes are used to generate pairs of corresponding subsets. The general principle for set generation, of which these three properties are

specializations, is: if graph G is isomorphic to graph G*, the subset of nodes of G exhibiting some property must correspond to that subset of nodes of G* exhibiting the same property. The second basic part of the algorithm, the partitioning procedure, is required whenever a number of pairs of corresponding sets have been generated. The purpose of the partitioning is to reduce the number of possible correspondents of each node of G to as few nodes of G* as possible (ideally to a single node).

The algorithm, as programmed for the IBM 7090, requires about 7,000 words of storage and can accommodate graphs with up to 200 nodes. In a limited test of the system, 50 compounds were chosen from Beilstein to form the library. Each compound had exactly 50 atoms and contained only the elements carbon, hydrogen, oxygen, and nitrogen; only six different molecular formulas were represented. The compounds were selected in this manner to utilize the full power of the algorithm frequently. All the compounds were reused as queries so that the algorithm was employed 1,275 times. The time to determine that two compounds were not isomorphic was, in 85 percent of the cases, less than 0.5 millisecond; the longest time was about 100 milliseconds. The time to compute a complete isomorphism ranged from 3 to 13 seconds; when all hydrogen atoms were removed from the compounds, the time was reduced to the range of 0.8 to 4 seconds, and no backtracking was required.[178]

Techniques and computer programs have been developed at Chemical Abstracts Service for generating a unique description for the two-dimensional projection of known chemical structures.[133] The third dimension is handled by addition of conventional stereochemical descriptors supplied by the chemist who prepares the structural diagram for input. The third dimension will be included directly in the graphic record to give a single, detailed, coherent record of each structure.

The structure description used in the CAS registration process is in the form of a connection table; within the computer the form is translated into a compact connection table. At first, this table is an unambiguous but nonunique machine representation, one of a family of unambiguous descriptions of the structure. The exact table selected for use in the Registry System is a member of this family, selected by further computer processing. The system is based on the premise that a unique structure will be stored once and only once, thus making the registry number a unique and unambiguous identification of a chemical substance.[148]

The computer program generates an invariant subset of the possible set of all tables describing a structure, orders the members of that subset, then selects the first member of the resulting list as the

unique table. Methods have been devised to reduce what would otherwise be an impractical, time-consuming task to one requiring a reasonable amount of time. The techniques were programmed for an IBM 1410 system; over 25,000 chemical structures from CAS files were processed as a test to demonstrate the economic advantages.[133]

A general technique proposed by Armitage and Lynch for the automatic determination of structural similarities among pairs of chemical compounds is based on the generation of fragments of each structure, starting with the individual atoms of each and, by concatenation, fragments of increasing size.[4] Each fragment generated comprises full information on the constituent atoms and the bonds connecting them. At each step in the iterative process, the fragments formed from one structure are compared with those from the other; non-common items are discarded, and growth continues only from those fragments common to both. The procedure is continued until the structural "highest common factor" has been determined. It is evident that if each common factor discovered is distinct and if the process is continued through even a relatively small number of iterations, the number of individual comparisons rapidly becomes very large, so the procedure is likely to be impractical, even if each comparison gives rise to only one highest common factor. An alternative approach to the repeated comparisons is to have all possible simple fragments generated for each structure; pairwise comparison then involves matching the organized lists and synthesis of the common factor only. While the generation of all fragments is costly in time and storage, the amount of computation in the individual pairwise comparison is greatly reduced. The advantages and disadvantages of this approach require further study.

Shell Development Co. reports a method for the reversible and unique encoding of a graph as a linear symbol string suitable for storage, classification, recognition, and retrieval purposes.[176] The graph is converted to a representative character string without loss of any topology, and the resulting character string is completely independent of the nontopological properties of the input representation. Transformation of the input form of a graph to a character string is linearization of the graph; transformation to a unique character string is standardization. An algorithm was developed to solve the problem for certain classes of graphs by constructing the standard adjacency-matrix representation of the graph. The goal for further work is a practical and automatic method, mathematically rigorous and embodied in a computer program for standardizing any graph of possible interest.

E. INTERCONVERSION OF STRUCTURE REPRESENTATIONS

A number of different linear notation systems, proposed or applied in operating situations, have been claimed to have the ability to describe uniquely and unambiguously any organic compound by a row of symbols consisting of letters, numbers, and punctuation marks such as are found on a typewriter. In any one of the proposed systems, there exists for each compound a single correct cipher called the canonical form. It is characteristic of different notation systems that different symbols are used to represent the same structural features and that different orders of precedence are prescribed for feature citation. As a consequence, an alphabetical listing of ciphers from one notation system will bring together compounds that would be scattered in an alphabetized list of ciphers from another system. Which compounds are better clustered and which are better scattered depends on the characteristics of the situation for which the system is proposed or operated.

Arguments have been advanced citing the advantages of one system over another. The substance of such arguments has been based mainly on the relative merits of the forms originally conceived for the various ciphers. Today, the ability of electronic computers to convert arrays of symbols to differently ordered but equivalent arrays makes it imperative to question some of the strongest traditional arguments. Hayward suggests that those who work with a given chemical notation system need no longer stay with ciphers solely in their original canonical form, but may elect to generate through machine programs a variety of equivalent forms tailored to serve specialized purposes.[80] Structural features implicit in a given form may be made explicit in another. A sequential cipher may have a hierarchical image. That aspect of structure emphasized in one hierarchical arrangement may be suppressed in another. A contracted form may be expanded. As Hayward puts it ". . . a whole battery of cipher forms, each unique and unambiguous within its own linguistic rules of order, may be generated for each chemical structure." Attempts have been made, also, to convert ciphers of one notation system into those of another. It remains to be determined whether the syntax of each system permits cipher interconversion to the same degree. The increased versatility and scope added to the use of ciphers through their manipulation by computers justifies a more penetrating study and evaluation of the chemical notation systems that have been proposed.

A notation in its original form may not be altogether suitable as a method of representing structural data for computer manipulation. However, the notation can be converted by computer into a connection table, a form that gives a precise, chemically descriptive rec-

ord suitable for all computer applications. In this section we review various efforts at converting coded representations from one form to another.

Conversion of WLN

Imperial Chemical Industries, Ltd.

An investigation conducted by Hyde and initiated by Imperial Chemical Industries, Ltd. (ICI), considered an atom-by-atom connectivity system based on mathematically derived matrices.[88] The investigation showed that this method was too cumbersome for the proposed system and that in many cases it obscured the familiar association of the atoms in organic structures. Further investigations have shown that the Wiswesser notation effectively describes the chemistry and the topology required for the system. A computer method has been devised for producing a matrix directly from the notation in a form suitable for search and correlation purposes. Hyde cites as reasons for use of the Wiswesser notation: it has overcome a number of the problems of an atom-by-atom system; it is canonical in the linear ordering of the notation symbols and this ordering has not destroyed the arrangement of the atoms in the molecule; it is concise because bonds and atoms have been compacted into one symbol, and, due to the linear arrangement, there is no need to state connectivity; finally, it has enriched certain atomic symbols to the point where their chemical significance and differences are clearly shown. Scrutinizing a molecule by computer becomes a much simpler task because the symbols act as a fragment screen.

The Wiswesser notation does not spell out every single atom in a molecule; therefore, it is necessary to generate from the notation all excluded atoms because these constitute nodes in any derived connectivity network. This condition is met mainly in ring systems. In earlier work, Wiswesser had been working on a system, "Dot Plot," for spelling out every node in the rings using the following symbols for ring carbon: L for $-CH_2-$, Y for $-\overset{|}{C}H-$, X for $-\overset{|}{C}-$, D for $-CH=$, and T for $-\overset{|}{C}=$. These letters had been carefully chosen not to interfere with existing symbols in the notation and could be used to expand the ring notation and to provide the nodes essential for a connectivity network.[211]

A program has been written that builds a connectivity matrix using both Wiswesser notation symbols and Wiswesser "Dot Plot" symbols, working character by character through the notation.[88] The connectivity matrix for a molecule is described, therefore, by a tape

record composed of three sections: units, connection transfers, and ring block. This form has the following advantages: (1) the units are readily accessible for screening purposes, (2) the terminal and branching units are recorded so that the network of the molecule can be rapidly reconstructed for generic and atom-by-atom searching, (3) the ring atoms are clearly identifiable as being in the same ring, (4) information on ring size is available, and (5) position of fusion is indicated.

The linear and numerical form of the compacted matrix lends itself to rapid screening and identification. The ICI staff have used this matrix to generate an open-ended fragment code[188] and to generate chemical structure display by the computer.[90] The display program, written in COBOL for the IBM 360 system, incorporates a large number of detailed statements for producing a conventional representation of a structural formula, omitting bonds in such groups as $-SO_2-$ and $-COO-$, for example, and arranging in a conventional order the atomic symbols that begin a structure, as in NO_2-. The plotting routine allows for eight possible tracking directions that can be applied either to chains or to rings and their substituents.

National Bureau of Standards

A project at the Center for Computer Sciences and Technology of the National Bureau of Standards was aimed at developing means of handling chemical structure information in such a way that several installations with different methods of representing structures and with different requirements on structure searching could supply information to a common system and call on it for structure searches.[187] The effort was concentrated almost completely on designing algorithms for transforming Wiswesser notations into connection tables.[61] In addition, routines were written to transform connection tables into sets of matrices suitable for searching by the Sussenguth technique (described earlier in this report) and to extract keys for sorting, consisting of an integral number of computer words.

Some machine use can be made of the sequential cipher directly for searching, but the chief utility is viewed as providing a convenient source of the structure information for other forms more suitable for most machine manipulations, including generic structure searching. Different features of the structure are given emphasis in the several forms, but each representation is unambiguous, that is, the complete structure diagram can be directly reconstructed from any one of them.[185] The project emphasized meeting chemical information searching and retrieval needs of the U.S. Patent Office, although

the results are believed to be potentially applicable to large-scale chemical systems in general.

University of Pennsylvania

An algorithmic analysis of the Wiswesser Line Notation undertaken at the University of Pennsylvania resulted in a program to decode from the WLN to a connection table.[103] The mechanical translation, so-called, derives from properly encoded notations based on the rules in Smith's draft of the system's manual.[167] Because of the distinction in rules for acyclic and cyclic compounds, the program was written as two separate subroutines, one for each type of compound. In addition, compounds containing both acyclic and cyclic portions required the writing of an executive program to select the relevant subroutine at the proper time and to assemble the results of the two programs.

Three general reasons are given for the problems encountered in completing the translation system: (1) certain WLN rules (or situations created by the lack of certain kinds of rules) are not amenable to algorithmic interpretation because the chemical knowledge and experience used by a chemist to resolve potential ambiguities cannot easily be formalized as an algorithm; (2) the use of contractions and multipliers in the notation adds greatly to the complexity of decoding because they represent a shorthand of the actual notation and the compound is adequately represented in the uncontracted and nonmultiplied form; and (3) the executive program (which controls the switching to cyclic or acyclic processing) needs an additional feature to enable processing of compounds with more than one cyclic nucleus.

The programs are written in FORTRAN IV and operate on the IBM 7040 and 360 series. The normal program output is a redundant connection table with no compression, printed in a tabular format. The flow charts used to implement the translation are contained in the second volume of the 1967 annual report of this project.[104]

The Dow Chemical Company

In order to run group or fragment searches (to locate compounds that have certain elements, or groups, or structural features in common), the staff of The Dow Chemical Company carried on some preliminary work on their file of Wiswesser notations.[105] They wrote an ALGOL program that will operate on such a file, arranged to produce on magnetic tape an inverted fragment-searching file. This program, "Fragmenter," analyzes the notation and produces a series of code words describing the individual features of the com-

pound. The number of such code words developed depends upon the complexity of the compound. Each code word contains in its lower-order digits the compound number, while the upper-order digits (the code) describe some features of the compound.

There was a great deal of redundancy, which had to be handled during the searching program. The computer sorts the code words in ascending numerical order, which places all like codes together; within any run of like code numbers, the compound numbers are also in ascending order. The files can be searched by a general searching program that compiles lists of compound numbers of those compounds having certain features in common. The process of fragmenting by computer is claimed to be so easy that fragment files of extensive details and coverage are feasible, can be refragmented periodically as ideas and interests change, and can be made available in different forms for the convenience of research groups of different orientations. Also, since the answers come back from the computer in the same arrangement and order as in the original inverted searching file, the new compound number list can be read back onto magnetic tape as an addition to the searching file. In those cases where a final detailed structure search is necessary to answer the proposed question, the preliminary use of a group search as a screening device will greatly reduce the number of compounds that must be searched in the more difficult fashion.

The Dow group also described algorithms that could be used either manually or by computer to translate a Wiswesser Line Notation into a tabular connectivity form, and to draw the picture or structure from the notation. Each of the forms has different properties and uses: structural diagrams are preferred by chemists, the connection table is fine for atom-by-atom searching, while the linear notation is useful for computer input and storage.[105]

Lynch

A notation using the symbols and syntax of the Wiswesser Line Notation and the ordering of a canonical connection table from which it can be derived is suggested by Lynch for use in mechanized systems.[119] The translation from connection tables into linear notations is a more complex task than the reverse translation. The solution suggested is the conversion of a notation file into connection tables, to produce the canonical forms, and then conversion back into synonyms of correctly ordered WLN's. A computer program, written in the list-processing language, SLIP, accomplishes this for acyclic structures. Extension of the procedure to cyclic structures is proposed.

Conversion of IUPAC System

Algorithms were developed at Chemical Abstracts Service for converting IUPAC notation ciphers to two types of matrices.[50] This development was accomplished in two stages: first, the development from the cipher of a "cipher matrix" and, second, the development of a machine matrix or "extended matrix," literally a fundamental chemical-structure language in which the hierarchical fragments of the cipher are obliterated. Thus, the extended matrix is one of a series of related random matrices and is independent of the characteristic features of any generating cipher. The first, or cipher matrix, is "cipher-directed"; the extended matrix need not be so directed, although for convenience it often is.

It was assumed that the proper input to any system dealing with chemical structure is a correctly drawn structural formula. This geometrical pattern, or ideograph, is converted by the process of ciphering to a linear algebraic form—the cipher. The deduction of such a cipher from the ideograph involves numbering the atoms of the latter according to a predetermined set of rules. This set of rules can best be applied by the trained scientist. An algorithm was constructed and programmed for checking routines and substructure searches, which accepted the cipher as input and computed from it the cipher matrix and the expanded matrix; the latter was used for mechanical computation of the molecular formula or for ascertaining the presence or absence of any substructure. The procedures, geometrical structure ⟶ cipher ⟶ cipher matrix, were performed in the first stage by scientifically trained personnel, and in the second stage, by computer. To decrease the use of skilled manpower needed in this sequence, it was necessary to replace the scientists used in the first stage by nonscientifically trained assistants and to use an algorithm to convert numbering applied randomly to cipher numbering.

This conversion system is not in current use at CAS.

F. CODING INORGANIC COMPOUNDS

The vast number of organic compounds represents a great challenge to chemical information services, whether conventionally oriented or computer-based. Certainly, the major amount of effort reported in the literature and reflected in the pages of this review has been devoted to coping with organic structural formulas. But the storage and retrieval of inorganic compounds has also received attention, and

rules have been developed for writing equivalents for conventional inorganic formulas.

Many inorganic compounds, salts, and salt-like compounds differ from typical organic compounds in being aggregates of ions rather than fixed arrangements of atoms in molecules. McDonnell and Pasternack described an initial attempt to develop a method for handling inorganic coordination compounds that should be as compatible as possible with existing practices in organic notation systems and that could be extended to other types of inorganic compounds.[126] Their approach is independent of any particular set of seniority rules, but, for purposes of illustration, the principles of the Hayward System for organic structures were utilized.

The class of compounds chosen for the development of their system was coordination complexes, which include discrete aggregates (cations, anions, and neutral molecules) rather than endless chains, lattices, or networks. The development of methods for uniquely and unambiguously describing attached group (ligand) positions in coordination compounds has become of considerable importance as a result of recent advances in inorganic and structural chemistry.[149]

Coordination centers (termed "Group M" atoms in this system) exhibit variable coordination numbers and symmetries. Unique and unambiguous representation of these complexes requires the development of a method for expressing not only the relation between each attached group and the coordination center but also the relations among and within the various ligands themselves. Furthermore, if several Group M atoms are present in a given aggregate, their relationships one to another must be described.[126]

A table of reference structures was prepared containing idealized representations of postulated structures and showing the highest-order axes of rotational symmetry and planes perpendicular to those axes in which ligand attachments lie. Each proposed coordination structure is assigned a two-character symbol: the coordination number of the fully coordinated Group M atom and an uppercase letter, the "symmetry designator," that refers to the particular geometric distribution of attachments. Furthermore, each position of ligand attachment is assigned a single lowercase letter, termed a "locant designator." The use of these locant designators in the actual ciphering of a given aggregate depends upon a set of arbitrary seniority rules. With the aid of such rules, the relative seniorities of ligands about each Group M atom are determined. Since several coordination centers may be present in one aggregate, the Group M atoms are assigned index numbers according to some arbitrary method, for example, by a set of rules similar to the Hayward rules for enumer-

ation of parent hydrocarbons. Group M atoms are ciphered first, ligands are ciphered next as independent entities, and bonds between Group M atoms are ciphered last. Various types of stereochemical information and enclosing marks are also introduced.

The approach outlined in this system permits the coding of coordination complexes in a linear cipher suitable for computer storage. Since more than one complete name could be assigned to each structure because of accidental differences in representations, an arbitrary set of seniority rules can be established to rotate the representations into particular orientations before locant designators are assigned.

The work of McDonnell and Pasternack proposes a method of ciphering coordination complexes using a set of idealized structures. However, many important properties (such as spectra, and magnetic susceptibility) of this class of compounds depend on the true symmetry of the species. Silverton and Pasternack suggest that it is appropriate, therefore, to include the true configuration in the store of information about the given substance.[163] The idealized structure can be represented in one file by the McDonnell-Pasternack cipher and the true configuration, if it differs from the ideal, and in a second file it is represented in terms of its point group notation. A deviation is termed major when it cannot be attributed to inaccuracies in atomic parameters. The estimated standard deviation serves as a criterion for classifying apparent deviations. When all deviations are minor, the structure is treated ideally.

The point group of the true configuration contains only those symmetry elements of the idealized configuration from which there are no major deviations. Therefore, the first step in determining the true configuration is to test the symmetry elements of the idealized configuration for deviations. Three types of symmetry elements are used in point-group notation: inversion center, rotation axis, and reflection plane (mirror plane). Combinations of these symmetry elements produce additional symmetry elements.

As noted elsewhere in this report, several areas of development of the Chemical Abstracts Service Registry System include addition of new classes of compounds, particularly inorganic and coordination compounds. Such compounds will be handled by modifying the conventions now used for organic compounds. Coordination compounds having six or fewer attachments to the central atom are being put into the present CAS Registry System. Reprogramming of the structure registry system for the IBM 360 computer will allow input of structures having as many as 32 attachments to any one atom.[3]

Less sophisticated methods for handling inorganic compounds are in operation while these more detailed techniques are being developed. For example, Kirschner and associates at Wayne State University have established a computer system to retrieve information on anticancer activity of coordination compounds.[98] The system uses keyword headings, making it possible to search for particular metal ions, ligands, functional groups, and so forth.

III

TOTAL SYSTEMS DEVELOPMENTS

*

There has been growing awareness of the need to consider as a whole all aspects of a chemical information handling system: collection of data, input processing, correlation and computation, manipulation, output production, and all other procedures involved in the performance of services. The data may come from various sources, need multiple manipulations, result in a number of final products, and so on. By considering all aspects of the total process, better design and more effective operations are possible. In this section, some particularly significant developments in the design of total systems are reviewed.

A. CHEMICAL ABSTRACTS SERVICE

The Chemical Abstracts Service (CAS) information-handling and publications operations are steadily being converted to computer-based procedures. A key point characterizing the CAS approach to large-scale information systems is the "single analysis/multiple use" concept.[46] In the development of these systems, there is no intention to eliminate printed publications. There is the aim to use the computer's capability of preparing both printed and machine-readable information services from the input of a single data stream and to eliminate duplicate intellectual effort in the analysis of the original material from which these services are prepared.

Currently, magnetic-tape records of four computer-based ser-

vices are available, some with corresponding printed records. Chemical Abstracts Service performs searches of these tapes on request, but subscribers are now encouraged to acquire the machine-readable files and to do their own searching.[94] Included in a subscription are: copies of the printed issues, search guides, and manuals for preparation of search profiles. Search programs (and associated documentation) written for minimum computer configuration will be provided as needed.

In the past decade the number of items abstracted in Chemical Abstracts (CA) more than doubled, and in 1966 about 243,000 bibliographic items were covered. By 1970 the bibliographic items will increase to about 360,000 per year, and in the decade 1967–1976 it is estimated that about 3.54 million abstracts will appear.[182]

The traditional system of abstracting, indexing, and publishing is a redundant process. Each of the steps requires some intellectual input and must be carried out by a member of the professional staff. On the clerical side of the operation, the principal redundancy is multiple "keyboarding"—the mechanical translation of information by typewriter, keypunch, monotype, and the like from one printed form to another. In the computer-based system, a single keyboarding will put on magnetic tape all data selected in a single intellectual analysis of the source documents, an analysis combining both abstracting and indexing. The computers can then produce title-alerting publications and bibliographies, material appropriate for special-subject alerting and retrieval publications, the contents and indexes for CA itself, and other desired information.

All these publications will be produced through a composition system operated by the computer, without further intellectual manipulations of the information and with little additional clerical effort. Equally important, the information will be recorded permanently in machine-searchable form for immediate and future reuse. In 1965 CAS began using a special 120-character IBM 1403 print chain that has the full Roman alphabet in uppercase and lowercase, 12 commonly used Greek letters, 14 special symbols, and on-line, superscript, and subscript numerals.

In September 1967, Rule reported the beginning of conversion of CAS composition to a modified IBM 2280 Film Recorder that can record all of the nearly 1,500 symbols required to compose the CA issues.[155] Characters are formed by programmed stroking of 35-mm film with an electron beam. Film output is converted to offset plates for conventional printing. The quality obtained through this process is excellent (equivalent to hot type), and composition proceeds at a rate of between 1,500 and 5,000 characters per second, depending upon the character range used and the printing quality desired. Con-

tinuing experiments demonstrate that molecular structural formulas can be composed on-line with the text, eliminating the need for artwork and film-stripping.

The total CAS system is developing rapidly. The computer-manipulable data store contains (as of the end of 1968) all bibliographic data processed for Chemical Titles (CT) since 1961 (730,000 titles) and, for Chemical-Biological Activities (CBAC), whole digests processed since 1965 (53,000 digests). The indexes of these two publications are also in the data store. In addition, all compounds indexed in CA since January 1965 have been entered in the Compound Registry System—a total of more than 1 million compounds with more than 1.25 million corresponding bibliographic references, to date. Previously unregistered compounds are being added at a rate of 5,000 to 7,000 per week. All titles and bibliographic data in CA issues, for 1968, numbering some 233,000 bibliographic items, as well as the 1968 issue and volume Formula Indices, are to become part of the machine record. In 1969 and 1970, respectively, the volume Author and Subject Indexes will be added, and, from 1972 on, the entire contents of current CA issues will go into the record.*

One of the control measures being instituted by Chemical Abstracts Service is the development of a computer-based inventory of abstracts, to provide better insight into, and control of, production operations.[17] The primary tasks of this system will be to control records of acquisition, assignment, and coverage, to detect duplicate records, to track material through the processing cycle, and to maintain production statistics.

CAS is thus in the midst of a step-by-step conversion of all operations to a computer basis. The approach has been to proceed in an orderly fashion from pilot-scale operation to full-scale production.[46] The introduction of new computer-based current-awareness services has been paced deliberately: Chemical Titles in 1961, Chemical-Biological Activities in 1965, Polymer Science and Technology in 1967, and in 1968, Basic Journal Abstracts containing the abstracts selected from 33 core journals. Each of these services has a corresponding magnetic-tape search service and is available for subscription in magnetic-tape form. The latest service, CA Condensates, provides a computer-searchable record of titles, bibliographic data, and keyword index phrases for all CA abstracts published since mid-1968.

As reported in Survey of Chemical Notation Systems, CAS started research in 1959 under the direction of G. Malcolm Dyson on the ap-

*Private communication from Russell J. Rowlett, Jr., Editor, Chemical Abstracts.

plication of modern data-processing techniques to the handling of chemical information. By 1964, this research, combined with earlier work performed at Du Pont, enabled CAS to embark on a five-year program to implement an operating computer-based system. The majority of the earlier work was in the area of chemical-compound searching because 85 percent of the entries in the CA Subject Index are related to chemical compounds or materials.[214] Furthermore, the structure area appeared more amenable to early solution than did the concepts area. The net result was the establishment in 1965 of the CAS Registry System.

The CAS Registry System operates by computer development of a unique connection table and computer assignment of a unique number, called a Registry Number, to each compound when its structure is first entered into the system. Whenever a compound that is already on file is entered, the previously assigned Registry Number is recovered automatically. Those compounds that are reported frequently in the literature can be arranged by the computer into an index of names and corresponding Registry Numbers. Instead of having to draw the structure for such compounds each time they occur, the analyst simply looks up the number, and the bibliographic and nomenclature files are updated on this basis. Such aids to indexing are referred to as Desktop Analysis Tools (DAT's). Associated with the DAT's is the proposed Registry Handbook, which, in its preliminary form, consists of microfilmed Registry sheets including the hand-drawn structural diagram, the molecular formula, and the Registry Number for each compound in the file, arranged in Number order.[113]

The system has three associated computer files: the Structure File, the Nomenclature File, and the Bibliography File. The Registry Number functions as a machine address within these associated files and links together all information about a given compound. For instance, it ties the information in the Nomenclature and Bibliographic Files to the structural description in the Structure File. It is the intention of CAS that when the system is completed, the user will be able to approach the system from any facet of the registered data and be able to retrieve desired information regardless of where it is filed.[3]

The CAS Registry is an integral part of the CAS index production system. The Nomenclature and Bibliography Files contain not only the CA Index names for a given compound but also all author names and synonyms so far encountered for the compound. Thus, an indexer may enter the name found in the original paper and retrieve the correct CA index name, the molecular formula, and the structure if needed. His time will be saved. He need draw the structure, calculate the formula, and develop the Index name only for those compounds

unavailable in the Registry. As the Registry grows, more compounds will be indexed without this repeated intellectual effort.

As entered into the computer, the description of the compound is unambiguous, but not unique. In the registration process, the input connection-table record is converted by program to an unambiguous, unique form. The unique forms of the tables are ordered in the file. Thus, the actual registration amounts to the merging of a list of additions with the Registry File.[113]

The Registry System is not simply an address-assigning operation. For example, it also functions in the over-all CAS system as a route of access to information associated with the compound. This point is illustrated by the use of the Registry System in the computer-based publication of Chemical-Biological Activities. In the CBAC computer files, biological test data are tied to the compound through the use of the Registry Number. The CBAC publication system provides not only a printed form but also a magnetic-tape form along with a searching program.[180]

The several areas of development for the Registry System include: addition of new classes of compounds, improvements in structure and nomenclature conventions, improved computer capabilities, improved structure registration, and new equipment and programs. The new classes of compounds are inorganic and coordination compounds, partially determined structures, and chemical elements. Inorganic compounds will be handled by modifying the structure conventions now used for organic compounds, thus retaining full structural specificity and the possibility of substructure searching. Structure conventions and input procedures have been developed for one group of partially defined structures, the definitive oligomers. This type of structure can be defined as a structure comprised of a single unit (the monomer) repeated a known number of times. The structure of the monomer is defined completely in terms of atoms and their connections. Finally, a program adjustment has been made to permit the registration of the elements and their isotopes.[3]

The CAS Substructure Search System retrieves compounds based on structural features. Computer programs can effect an atom-by-atom, bond-by-bond comparison of the structural features sought with the structures of the compounds on file; however, before this step is taken, many irrelevant structures can be "screened out." Information retrieved is in the form of the structures, names, bibliographic references, and Registry Numbers of all compounds on file that contain the specified features. Because the Registry Number is an unambiguous link for organizing all compound-related data in the CAS files, as described above, the path is also open for correlative searches relating structural characteristics to other pa-

rameters, such as physical and chemical properties, biological activities, and applications.

Searches on individual structure records are time-consuming and expensive. At least three techniques can be applied to increase the speed of search: screens, search strategy, and file organization.

Screens are characteristics of the compound, stored with the structure record or generated during search. The query is analyzed for the same features. If the record of these features in the structure files does not correspond to the inclusive or exclusive conditions of the query, the compound is not examined in the atom-by-atom search. Several screens are now operational at CAS.[102] These include total atom count, molecular formula, ring-closure count, bond types, occurrences of atom–bond–atom fragments, and more complex fragments. The efficiency of a given screen depends not only on the characteristics of the screen and of the compounds on file but also on the characteristics of the questions. Therefore, proper development and evaluation of screens requires a flow of "real" questions, and, in full operation, adjustments in the screens will be made as the question pattern changes.[59]

Modifying the statement of the query means that, for example, a structure's less common characteristics—atoms, bonds, groups, etc.—are placed among the first parameters of the search query, so that the presence or absence of the characteristic in the stored record can be determined as early as possible. File organization, on the other hand, presumes that it is possible to segment the collection according to criteria that reflect the nature of the queries. Thus, if more searches should include noncarbon atoms, there might be value in segregating the compounds that contain only carbon and hydrogen. A decision on file organization must await analysis of a broad spectrum of actual questions. Lynch has reported on research in process at the University of Sheffield "with a view to determining screens for substructure search which would have an even distribution through the file."[169]

Another phase of the CAS research effort has as its objective to develop computer-based methods to store, manipulate, and retrieve nonstructural, nonnumeric information.[59] The work is based on the use of words to describe such information, but nonverbal techniques are not ruled out for future consideration. The specific goal of early studies in this area was to extract from the CA indexes and indexing methods the essential parameters and relationships present therein. These findings were to form a basis from which to design and test computer-based searching tools for conceptual information. Compatibility with the Registry System and the substructure searching capability was, of course, required.

An interesting though unrelated scheme was outlined by Lynch[118] for automatically compiling and editing subject indexes by transforming descriptive phrases with regular structure and vocabulary. These transformations, based on the formal structure of language, were shown to be well-suited to computer manipulations.

Information in CA subject and molecular formula indexes should be easier to find as a result of a number of changes in indexing practice incorporated into the indexes now being issued. The Hetero-Atom-in-Context (HAIC) Index, a new type of molecular formula listing, has been added to the Formula volume. The computer-produced HAIC Index orders molecular formulas according to the noncarbon, nonhydrogen elements they contain and highlights these hetero elements within the context of the complete formula. The HAIC Index is a by-product of the conversion of the CA formula indexes to computer-based production methods.[93]

Statistics on the Registry System files present interesting information on the chemical characteristics of chemical compounds.[16] Since the Registry file is one of the world's largest organized bodies of chemical compound-oriented information, this information is perhaps available from no other source. Because of the large number of compounds registered and the variety of sources from which they have come, the file is broadly representative of all organic chemistry and contains relatively little bias.[115]

Two immediately apparent problems face the CAS research effort.[45] The first is that of size: the files will get exceptionally large as information from early volumes of CA is incorporated (about 3 million structures, 4–5 million names, and over 10 million references), and they will continue to grow indefinitely. In order to achieve reasonable economies of operation, it is necessary to limit processing to comparatively infrequent batch runs. The second problem is that of time: the process of retrieving information is limited by the availability of the information. Furthermore, the time required for chemists to initiate an inquiry and to refine it over a series of several retrieval runs would drastically lengthen the elapsed time between question conception and final answer.

A possible answer to both of these limitations is the development of a direct-access processing system and associated real-time retrieval operation.[215] However, the size of the store of chemical information provides a formidable challenge in the development of access storage-and-search techniques. By 1973, over 3 million compounds will be in the store if the necessary funds become available. Also, by 1973, abstracts will be accumulating at the rate of almost 400,000 per year, and index entries will accumulate at a rate of 3 million per year. The files must be organized to obtain the best bal-

ance of storage costs and access time for a variety of access approaches and search strategies. There are suitable techniques (for example, list structuring) for direct-access handling of stores limited to thousands of items. However, when the store grows to millions, these approaches become costly due to the excessive length of chains. CAS is investigating alternative approaches such as file partitioning and compact storage of efficient screens for rapidly scanning large blocks of information. It is too early to be able to determine the effectiveness of these approaches.[46]

The Conference on Mechanical Processing of Chemical Information, held at Airlie House in Warrenton, Virginia, on March 5–8, 1964, was summarized by Waldo.[38] Its purpose was, in addition to studying means of achieving an integrated mechanized system, "to make recommendations that would be useful . . . in determining . . . the financial support needed to bring [the system] into being." The National Research Council Committee on Modern Methods of Handling Chemical Information (precursor of the Committee on Chemical Information), which organized the conference at the request of the National Science Foundation, recommended that the American Chemical Society and Chemical Abstracts Service assume a large role in this development, with the responsibility implied therein. In this connection, CAS and the parent American Chemical Society are experimenting to develop techniques for mechanized information interchange on a large scale.

One form of interchange is the exchange of abstracts between primary journals and secondary services. Common standards have been adopted for the ACS primary journals and Chemical Abstracts, permitting the use of ACS abstracts in CA with minimum additional editing. CAS/ACS cooperative experience, with introduction of CAS Registry Numbers in the Journal of Organic Chemistry and the CAS-prepared indexes for the Industrial Engineering Chemistry Quarterly, has demonstrated that the approach is sound. The significance of this experiment is increased by the ACS conversion of some of its primary journals to computer-based publishing. Chemical Abstracts Service and ACS are developing a compatible set of character-representation standards, file formats, and field-content and identification standards that will permit direct interface by magnetic tapes produced as co-products with publications.

Another form of cooperative interchange under way is the linking of systems where there are common information requirements. For example, structural information on compounds of common interest to CAS, the National Library of Medicine, and the Food and Drug Administration is being routinely processed in the CAS Registry System. Through these cooperative exchanges, the data bases of

CAS and other processors are becoming compatible. Through the Registry link among the various systems, it is possible to approach the store of any compatible system and to retrieve compound-oriented information from any of the other systems.

B. WALTER REED ARMY INSTITUTE OF RESEARCH

Over one hundred military and civilian research laboratories are engaged in the various scientific aspects of militarily important programs in the development of drugs for use as antimalarials, antiradiation agents, and antischistosomal agents. Reports are distributed to nearly four hundred individuals or institutions submitting chemicals for testing. Central coordination of the entire effort is accomplished through the Division of Medicinal Chemistry of the Walter Reed Army Institute of Research (WRAIR).[95] This central agency collects and analyzes all raw chemical and biological data arising in the various laboratories and generates reports, so that the total program may function in an efficient and effective manner.

Because of the large quantities of data, computers have become vital as a link in the collection and dissemination of information. In the first 18 months of the antimalarial program, 75,000 compounds were accessioned and distributed for approximately 200,000 biological tests at the primary screening level. The spectrum of testing and evaluation beyond the screening level is rapidly expanding as the leads developed in the first year of the program are being exploited.

Computer-tape systems are used to facilitate the accession of new compounds, the shipment of chemicals, the generation of test reports to submitters of compounds, the control of chemical inventory, and the generation of formal progress reports. In general, rather conventional methods are used to accomplish these functions.

Biological data are stored on magnetic tape; cards, paper tape, and mark-sense forms are presently used as input media. Chemical structure file maintenance and chemical search are also accomplished on magnetic tape. The greatest volume of biological data generated in the laboratory comes from the primary screening tests, the results of which are easily adapted to storage on computer tape. The repetitive data generated in screening tests may be conveniently entered on 80-column cards in fixed-field format. A few relatively simple calculations at the input stage are usually all that is required to reduce the raw data to manageable form and to flag test results that suggest significant biological activity.[96]

Five problem areas may be identified in the over-all functioning of the system: input costs, interpretation of input and file maintenance, synonym files, sorting and questions, and output costs.

Jacobus states that low-cost input has been achieved by the use of the chemical typewriter, discussed again later in this report. This machine, by virtue of recording coordinates, provides unique digital locations for the characters typed and, therefore, digitizes a chemical structure. The typewriters used, from the Mergenthaler Company, have been relatively satisfactory. Typists are producing approximately 1,000 correct structures per week or 50,000 per machine per year per shift at a cost of approximately $0.10 per structure.* Overhead figures vary according to the scheme of operation, raising the cost to between $0.15 and $0.25 per structure. Of the total input, approximately 15 percent of the structures are rejected because of typing errors. One and one half percent of rejected material is rejected by the system because of chemistry. For the typing of a structure, the chemical diagram must first be made available. The conversion of a name to a chemical diagram may cost between $0.50 and $1.

The tape from the chemical typewriter must be interpreted in order to produce a conventional connection table or matrix suitable for machine processing. The programs for converting the input to connection tables operate as two relatively independent functions. First, the program must reconstruct the structural formula on the basis of the data contained in the digital output from the typewriter. Second, the program must analyze the structural formula to determine the actual network for conversion to a machine-oriented format that later parts of the system can process by techniques of network analysis.

Reconstruction of the structural formula involves three stages. The first stage unpacks input, checks typewriter codes for validity and correct parity, and evaluates and converts coordinates. Validity and parity are checked by means of a table, and the same table gives the numeric value of codes used as coordinates. The second stage loads input characters into a matrix in accordance with the coordinates applying to them, making erasures and corrections as required. Input codes are translated at this stage into the characters they represent, taking account of case shifts. The third stage scans the completed page matrix line-by-line and writes the intermediate output tape, dividing the page into individual compounds.

The output of this first phase is a conventional structure or connection table similar to those used as input in other systems. Screen bits are computed and assigned to each molecular record. A screen bit is a yes–no indication of the presence or absence of a previously specified substructure, such as a carboxyl group. The screens, however, may relate to relatively exotic functions. They also become a

*Cost figures are to be viewed with caution, as stated in the Preface.

permanent part of the molecular record, to be used as an extension of the molecular formula file.

In processing the structural formula, then, the final result is required to be a connection table (usually referred to as the "Struc Table") listing the atoms of the compound in arbitrary order. The atoms are numbered sequentially, and for each atom there is a list of the atoms to which it is bonded, along with the type of the bond. If the structure is given in complete detail, the processing is relatively simple. If the structure has been compressed to strings of element symbols and auxiliary characters, additional processing is necessary. Two basic approaches to the problem of interpreting such strings were considered.[49] The first approach would be to analyze strings entirely by program logic, taking each element symbol separately and inferring the relation of its atoms to the other atoms in the string. In terms of an analogy with natural languages, this procedure corresponds with interpreting a sentence word by word, allowing for all the changes in meaning of a given word that can be produced by changes in context. To follow the linguistic analogy, an alternative approach would be to analyze the sentence in terms of phrases instead of individual words. Chemically, this approach would involve defining a set of glyphs (groups of element symbols and auxiliary characters), each of which would be defined as representing one and only one arrangement of atoms and bonds. The final decision was in favor of using program logic exclusively (with the exception of three glyphs that have been defined to deal with unique situations).

Another problem appears to be that of ambiguity. The present solution is to issue warnings alerting the chemist to check the interpretation of the program.

A permanent accession number cannot be assigned to a compound by the system until the compound has passed through the previous programs, unless a manual system is used in advance of the mechanical system. The assignment of an accession number means that all the previously available less favored numbers must be changed or that a synonym file must be maintained so that access from the preferred accession number to the records under the old accession numbers may be established.

The chemical diagram output must be combined with biological or other data for it to have meaning. While this combination has been achieved by hand for many years, primarily by copy techniques, and while it has been achieved in a digital manner by Waldo, by Rice, and by others, it must be emphasized that large mechanical systems must achieve the combined print mechanically. Because of the synonym problem and because the chemical order is different from the accession-number order, much sorting and merging are required to achieve this combined print.

Computer programming accomplishes the merging of alphanumeric biological data of mixed types with the output tapes from chemical search of the structure file. Chemical accession numbers, only, are extracted and written on a separate tape, and the number tape is then matched to the numeric synonym file, enabling the compound to be found in any of the biological files, regardless of how many different identification numbers it may have. The final tape is then put on the drum printer, listing the compounds in the form of structural diagrams, followed by the associated biological data.

The last critical question is the development of low-cost output. This system used the Data Products drum printer for easy engraving of changes to accomodate the changes in chemical fonts used in the development of the typewriter input system. The printer delivers 600 half lines per minute when printing chemical diagrams and 600 full lines per minute when printing biological data.

In describing this high-volume operational system handling chemistry and biology, the WRAIR staff members emphasize the chemical problems as well as the magnitude of the conventional data problems. A successful system must be supported by experts in systems analysis, programming, and data processing, as well as by mathematicians experienced in topology and information theory. This system was stimulated in its growth and reliability by the willingness and need to depend on it for the control of a large test program in medicinal chemistry.

C. CHEMICAL INFORMATION AND DATA SYSTEM

The Army's program for a general chemical structure information system attacks the problems of chemical information and data handling on a broad front. The ultimate goal of the Chemical Information and Data System (CIDS) is the development of a network of specialized chemical information centers, tied together by an effective communication system that would enable a requester to assemble chemical information and data relevant to a specific or general query.[190]

The several phases of the CIDS program, established officially by the Army Research Office in 1963, included the following tasks: (1) development and implementation of a plan of action for the system, including determination of needs and total resource requirements, information and data systems design, and pilot tests for the acquisition, storage, processing, and dissemination of data; (2) investigation of the communications requirements for the system; (3) evaluation and development of computer-programming techniques and related mathematical models required for establishing, operating,

and improving the CIDS; and (4) assembling an inventory of chemical information and data holdings among the participating agencies, with recommendations for the phasing of the holdings into the information system, and concurrently, a determination of each agency's chemical information and data requirements.

Among the fundamental criteria to be met by an effective operational CIDS are the following: (1) the system should provide the user with either the specific information required or references to sources where the information can be obtained, in a time compatible with the user's needs; (2) it should be able to answer many types of queries beyond those based solely on molecular structure or chemical nomenclature; (3) it should take into account that many people other than chemists need and use chemical information; (4) it should allow the user to present his queries in his own language and not require him to conform to rigid query formats; (5) it should utilize local and specialized information centers now in existence at various places within the Army research and development complex, referring queries, as required, to the appropriate authoritative sources; and (6) it should be in touch with other governmental agencies, industrial organizations, and academic institutions generating or compiling relevant information so that these, too, can be used as referral centers.[192]

The types of information to be fed into the system, maintained in the central files, and searched include: registry number; chemical structure, probably as a listing of atoms (nodes) and bonds (connectors); molecular formula; bibliographical citations; nomenclature, including chemical names, linear notation, trade names, etc.; location of data files information; kinds of data and information available at each location; and security classification and restrictions on release.[200]

Collectively, the kinds of data to be recorded should provide the history of an entry: occurrence, isolation, synthesis, manufacture, purity, physical properties, chemical properties, biological activities, assay methods, applications, and patents. Negative as well as positive data should be available, such as an unsuccessful method of synthesis, a demonstrated lack of a suspected biological activity, or a demonstrated inability or inferiority to function satisfactorily in a proposed application. Experience indicates that only a small percentage of the queries will require immediate answers; it appears generally acceptable if answers are supplied within a day or two, or up to a week or two, depending on the nature of the query. It is envisioned, therefore, that the system will operate largely in the batch-processing mode. However, provision may be made for real-time operation as required.[194]

In the operation of CIDS, a technical information scientist would first examine the user's query and, if necessary, communicate directly with the user for any clarification required. He would then convert the query by selecting the appropriate command structure (substructure search, molecular-formula search, key-term search, nomenclature search, routed search to Technical Information Center, key-term update, nomenclature update, bibliographic update, etc.), adding the necessary control information (priority, output format, etc.), and transmitting the arguments of the query to those acceptable by the system, through either the printed-term catalogs and thesauri or directly via on-line console.

Different searching techniques, as they apply to the CIDS requirements, have been considered.[193] They include: (a) the molecular-formula and atom-by-atom search, which involves comparing the node-connector tables of all compounds in the file responding to the molecular formula search with a single generic node-connector table; (b) the molecular-formula and structural-fragments search, determined to be impractical in this case; (c) the molecular-formula, structural-fragments, and descriptive-term search, in which compounds identified as having the correct molecular formula are screened against pertinent structural fragments and against general descriptive terms, a technique requiring but a small fraction of the number of searches required by the first technique; and (d) the molecular-formula and Wiswesser-notation search, predicated on the assumption that the notation discriminates satisfactorily among chemical structures corresponding to the same molecular formula, thus eliminating the need for structural fragments and structurally descriptive terms.

Several organizations are collaborating in the design and development of the Army CIDS. In this connection, part of Frankford Arsenal's (Philadelphia) project responsibility has been to survey the requirements of chemists and other scientists and engineers for chemical information.[165] At the University of Pennsylvania an experimental system is under development in two phases, operating in a real-time or batched mode.[109] The first phase resulted in a system utilizing the IBM 7040 central processor and IBM 1301 disc-storage file, which can be queried from a single remote teletypewriter. An expansion in both the number of remote-inquiry stations and query versatility was planned for the second phase. For retrieval, the first phase utilized molecular-formula, structural-fragment, and descriptive-key screening, with atom-by-atom searching. The second phase would permit querying by certain kinds of chemical nomenclature, including linear notations.

The central file under Phase I of the University of Pennsylvania system consists of three subfiles: Subfile 1, stored in the magnetic-disc memory and responsive to the real-time molecular-formula, structural-formula, and descriptive-key queries; Subfile 2, stored on magnetic tape and containing all chemical and nonchemical nomenclature associated with the compounds; and Subfile 3, containing the bibliography, sequenced by registry number. Subfile 2 could be printed periodically as a catalog for manual retrieval on the basis of nomenclature.

The most basic aspect of the system, differentiating it from many other automated retrieval systems and conveying a real-time search capability, is its use of the list-structuring programming technique.[111] Its effectiveness is determined largely by the character and number of keys in the system. These keys may be any that are used for record retrieval, such as structural fragments, molecular-formula ranges, and descriptors that indicate applications or properties of the compounds. The desired goal is a large number of well-discriminating keys that partition the file into lists of practical size, say, of not more than 2 or 3 percent of the file.

The initial system for experimentation contained about 230,000 chemical compounds from the Toxicological Information Center file, the Chemical-Biological Coordination Center file, and the Chemical Abstracts Service file.[197] The system uses structural fragments but has the added capability of computerizing the assignment of fragments both to compounds being registered into the file and to compounds embodied in queries addressed to it. The interpretive form of the structural-fragment keys is a computer representation of the node-connector table. Just as the computer search system carried the connection tables instead of the actual structural formulas, so it carries the structural fragments (which comprise the structural formulas) encoded as search keys.

The staff at the National Bureau of Standards also contributed to the development of a versatile system for handling chemical structure information.[187] A file structure combining list-processing concepts (for handling variable-length information records) with standard serial record arrangements (for identification information) was designed for a large chemical information system, and included both well-structured and unstructured information.[184] Representative data inputs from Army files were examined for manipulation in the system.[2] Much of the effort went into designing algorithms for transforming Wiswesser notations into connection tables, as described in a preceding section of this report.

D. CHEMICAL INFORMATION PROGRAM

In the introduction to this report, a federal interagency Chemical Information Program (CIP) was described, intended to provide support for the development of a national system. The focal point for the research effort under this program has been Chemical Abstracts Service, but other requirements and approaches are also being taken into account. Thus CIDS, which was described in the preceding section, may be considered part of the total program. Other research tasks are being conducted in close cooperation with CAS, under the coordination of the National Science Foundation.

At the Moore School of the University of Pennsylvania, Lefkovitz is studying the impact of third-generation computer equipment on the design of a large-scale automated chemical information system.[110] A characteristic feature of such a system is the large number of chemical structures to be searched. The costs of retrieval on a serial-search basis would be prohibitive. Therefore, screens, or retrieval keys, are assigned to each compound, analogous to those assigned to a document in a document-retrieval system. Thus, a rapid screening decision can be made on a large percentage of the file.

The new generation of computer hardware provides direct-access storage for several hundred million characters of information. If the total file could be stored in such a medium, and if there were an extremely efficient set of retrieval keys that would partition the file for a given query into a relatively small sub-file, then a system could be designed that could search the large file on a random-access basis in real time more economically than systems presently searching much smaller files on a batched basis.

Another feature of third-generation hardware is that a relatively large number of remote consoles can simultaneously communicate with the central processor during search. Still another feature is the availability of a wide range of input and output devices. The cathode ray tube (CRT), for example, can be used as an output medium, especially in a real-time system, for the display of structural formulas as well as for the display of textual information.

At present there appear to be two mutually exclusive schools of thought on the development of automatic chemical information systems. One is centered on the use of a connection table (CT) as the representation of chemical structures, the other on the line notation (LN) as the representation. Although no system combines the use of CT and LN in such a way that each representation is allocated those functions to which it is particularly well suited, such a hybrid approach is appropriate and technically feasible.

Comparison of CT and LN system functions has led to such a proposed hybrid configuration.[110] Another coded representation of the structural formula that was developed stands in a one-to-one relationship with the connection table, therefore retaining all of its search and registration properties; but at the same time, it is as concise for digital storage as a line notation. This coded representation, called a Mechanical Chemical Code (MCC) represents a synthesis of ideas contained in existing line notations, and its form is based upon a code suggested by Hiz. It functions completely internally to the machine. It is designed to optimize machine functions, and its rules of formation, whether for the purpose of encoding to or decoding from connection tables or structural formulas, need never be the concern of a chemist, but should be completely and readily specifiable in terms of a computer algorithm.

To test these proposals on a sufficiently large scale, a Monitor system was constructed to enable the study of the file organization techniques through computer-generated printed indexes.[112] The simulated system, programmed in FORTRAN, is based upon computer storage of the CT in a highly concise linear notational format (the MCC just described) and upon retrieval screens derived in terms of this notation. Substructure search is accomplished by the screening system which possesses the characteristics of completely automated and economical screen assignment, ease of programming, and both coarse and fine screening. A sample of 4,000 compounds from the CAS Registry System was manually encoded into the Monitor system for the reported testing.

Another phase of the CIP is under way at the University of Pennsylvania Department of Linguistics. Munz and Weaver are studying the fundamentals of chemical structure representations. Munz reported on the problem of recognition of chemical rings[136] and on linguistics and information concerning chemical structures.[139] Weaver developed a mathematical model for chemical cipher systems[203] and, in a related study, Munz developed a formal evaluation of a simple chemical cipher system.[138] Both of these reports demonstrate the relation between this work and that of Hiz and Eisman.

Still another study supported by the CIP was conducted at the National Bureau of Standards and resulted in the design of the Solid System, which can be applied in principle to any body of information but which was demonstrated by using chemical structural information.[48] It is proposed that once information has been entered into a computer system, translation to some "Computer Oriented Representation System" (CORS) can occur, and the operations associated with storage, updating, and retrieval can be performed by manipulating these Com-

puter Oriented Representations. Integral parts of the proposed Solid System are the fully automatic Numeric, Alphanumeric, and Binary Compressors, which compress information entering the system before it is stored in slow memory. In both storage and retrieval operations, these compressors expand the compressed information on an item-by-item basis as the system requires it.

IV

CODING OF CHEMICAL REACTIONS

*

Chemical reactions, like chemical structures, require codification. Description of changes accompanying reaction is one of the sets of information chemists need to carry on the development of chemical theory and practice.

A. INTERNATIONAL PATENT INSTITUTE

In the novelty searches performed by patent offices, it is often necessary to retrieve documents relating to a typical combination of unit processes. Because of the large number of possible combinations, it is sometimes difficult to classify such documents systematically. To overcome this difficulty and to facilitate the novelty searches, a study was undertaken at the International Patent Institute to investigate the possibility of coding and recording flow sheets of such combinations for mechanical searching.[47] Although the system was developed for the field of hydrocarbons and petroleum chemicals, it is conjectured by the authors that it can be adapted to other fields in which the sequence of operations is important. In the petroleum industry, hydrocarbon fractions are converted or purified and separated or mixed in order to prepare products with the desired characteristics.

The normal zones of columns 20–57 of punched cards are used in a matrix arrangement in which vertical sections represent unit processes, and horizontal zones represent unit operations; perforations

at the intersection of a section and a zone represent the flows between these unit processes and operations. Searches relating to combinations of unit processes or unit operations and other data recorded on the cards may be performed by a sorter. A particular combination of such unit processes or unit operations may be recorded on a large number of positions on the card. The searches have to be performed in such a way that the number of passes is reduced. Using a sorter with a multicolumn selection device, or even more complex machines, will increase the efficiency of searching.

B. VLEDUTS

Vleduts has described a system for identifying chemical structure changes in reactions.[20] Bonds and rings created and broken are listed in cipher form. Certain special symbols are introduced. For example:

Δ = the part of the code characterizing bonds being created

∇ = the part of the code characterizing bonds being destroyed

$\Diamond$ = information about the closing and breaking of the rings in the course of a characteristic reaction

Numbers and symbols written after the circle signify the number of members in the ring and the hetero atoms in it. For instance,

$\Diamond$ 6, N

signifies that the ring involved contains 6 atoms including nitrogen. Thus, according to Vleduts, it should be possible to use a logico-information machine for automatic compilation of an index of organic reactions.

C. GREMAS

Documentation of organic chemical reactions with the GREMAS system has been described.[64] In analyzing a chemical reaction for storage, those parts of the molecule that are involved in the reaction are recognized. The parts involved are mentioned in their initial and final states. The same faceted classification numbers are used that were developed in analyzing the whole structural formula.

D. COMPUTER ANALYSIS OF STRUCTURE CHANGES

Armitage and Lynch have described the application of their system for the detection of similarities among chemical compounds to the identification of structure changes in the reactions of acyclic compounds.[5] Their approach simulates some of the mental processes a chemist employs when he examines the equation for a reaction and deduces the nature of the changes taking place. The chemist, in scanning the equation, identifies the common features on either side as a preliminary to pinpointing the differences, thereby identifying the site and the nature of the change. Thus, a search for similarities is the first step in the search for differences. The stepwise generation of fragments from each structure is considered in turn. The smallest fragments are the atoms of the structure, the next are the atom-bond-atom pairs, etc. At each stage, the fragments of one structure are compared with those from the second, and vice versa, and only those common to both are considered for further growth. The process is continued, the fragments growing in size at each step, until no larger common fragments can be formed.

The comparison of fragments, at each stage in the procedure, must be performed many times during the analysis; it is essential, therefore, that it be conducted as efficiently as possible. It is, in fact, a search for identity. A general program for a search for identity makes considerable demands on computing time. For this reason, the fragments generated have been limited to simple structures for which canonical forms can be generated quickly and efficiently. The simple fragments are compared in canonical form and are subsequently built up to synthesize the actual common fragments, which may be complex and highly branched.

Initial effort has been concentrated on acyclic structures, and therefore on reactions in which both reactants and products are acyclic, since the acyclic case presents fewer complexities than when rings are also present. Programs have been completed and tested that analyze pairs of structures in terms of simple fragments and print out the results of the analysis. It should be noted that although reference has been made only to pairs of structures, reactions in which more than one reactant or product molecule take part can also be handled by this technique; the sets of compounds on each side of the equation are treated as disconnected graphs, the nodes being numbered by successive integers.

The concept of similarity among sets of chemical structures has far-reaching implications, not only in the analysis of chemical reactions, but in many other areas involving chemical structural information, including procedures that chemists intuitively use whenever they

survey chemical structures with a view toward relating structure and activities of various kinds, including chemical reactivities, physical properties, and biological effects.

E. REACTANT INDEX WITH WISWESSER LINE NOTATIONS

The concept of permuted Wiswesser line notations described earlier in this report has been extended to develop a reactant index.[67] The line notations of the chemical reactants were added to the punched cards containing the notations of the reaction products. Space is also available for indicating the catalyst and reaction conditions. The new index has enhanced the utility of the over-all program since it now allows the user to locate rapidly all products from specific or similar starting materials.

F. REACTIONES ORGANICAE

An index of chemical reactions, called Reactiones Organicae, published by Georg Thieme Verlag, Stuttgart, includes punched cards; a handbook with an introduction containing the necessary rules; two indices, one of initial and final groups (classed by chemical constitutions and by empirical formulas) and one of reaction conditions; and selection devices to manipulate the internally-slotted cards.[216] Each card describes one reaction; the upper half contains the natural-language text, and the lower half comprises the punched area. On the reverse side of the card are the name of the reaction product, the procedure, and other compounds that can be synthetized by the same procedure, along with their yields.

The coding of this data system corresponds to the various aspects of reactions: reaction centers, reaction conditions, and reaction products. The reaction centers of the initial compounds and those of the final products are recorded as individual formulas and by the generic chemical classes to which they belong. Each reaction is defined by a single group of physical and chemical reaction conditions and is recorded by a decimal code. The characteristic features of the molecular skeletons and the functional groups of the reaction products are recorded by direct code.

V

PROPERTIES INFORMATION

*

Effort has been devoted in many organizations to developing a computer-based system for the storage, manipulation, and retrieval of data on the physical properties of chemical substances. Such projects are outside the scope of this report. However, the following projects concern the estimation of physical properties from chemical information.

A. AMERICAN INSTITUTE OF CHEMICAL ENGINEERS

A project sponsored by the American Institute of Chemical Engineers and supported by industry has developed generalized computer programs for the estimation of physical properties of chemical compounds.[127,128] Given some information about a chemical, such as its structure, normal boiling point, and molecular weight, the program will estimate values for other desired properties at specified conditions. In general, the more data that can be supplied, the better will be the program's performance in estimating other properties.

Required as input are descriptors that give a coded description of the compound, including information on whether the chemical is an element, a hydrocarbon, or an inorganic compound; whether its molecules are monatomic or diatomic; whether it is a straight-chain or branched-chain paraffin; whether it contains an aromatic ring; and other factors. Such information is necessary since most of the estimation techniques are limited to certain classes of compounds. An

estimation technique included in the computer system is called a transfer function.[140] A transfer-function table in the program lists all the transfer functions by their output variables together with a list of the input variables required by each.

The first step in the program consists of the "road-mapping" routine. The second step consists of the "optimum-route-selection" routine. Both of these routines operate as organized searches. Searching is required in the road-mapping routine because it is not feasible to store a list of all such road maps in advance; the number of different combinations of known and requested properties is much too large. In the optimum-route-selection routine, a search is required because the accuracy of each route cannot be estimated until it is applied to the particular problem.

B. THE DOW CHEMICAL COMPANY

The estimation of physical properties by the general method of group contributions has been studies at Dow.[27] One of the purposes was to demonstrate the possibility of machine decoding of linear notations, which is particularly useful for the estimation of physical properties. What is desired is a simple input from which the computer itself will determine the individual groups required for the correlation. An essentially linear translation or decoding scheme is used. After the input data have been read, the number of symbols in the notation string is counted. Then the string is decoded by counting the main carbon atoms, side carbon atoms, all single groups, triple bonds, and double bonds. A feature that is unique to the project is the calculation of the molecular weight, the Lydersen constants, and the critical constants. Symbol counting is relatively easy since only a single space is legal in the Wiswesser notation used in this study. Hence, a double space is the end of the notation.

VI

MACHINE HANDLING OF NOMENCLATURE

*

In the Introduction to this review, it was noted that systematic nomenclature is an aid in identifying chemical compounds, and is suitable for indexing when it provides a specific address for a compound. At the same time, names provide a poor basis for classification or grouping of like compounds. The subject of chemical nomenclature per se is outside the scope of this report. However, since nomenclature plays an important part in the design and operation of systems for handling chemical structure information, studies concerned with its manipulation will be briefly reviewed.

For one thing, verbal and written communications are required by persons other than chemists. Such persons are not inclined to use graphic formulas and they must refer to materials by name. At least temporarily, they must rely on scientific names, until trade or generic names come into being.[123] The synonym problem which arises is an important factor in the design of systems for handling chemical structure information.

Two of the procedures used to manipulate nomenclature are name-editing programs and nomenclature diagnostic listings. Other techniques include computer translation to structural representations, computer generation of names, and computer retrieval of index nomenclature by name match. All of these procedures are aimed at extending the capabilities of systems for storing and retrieving chemical structure information.

Although routine editing of chemical nomenclature is an absolute necessity in a chemical information system, it is also costly and

time consuming when performed manually.[3] Park, at Chemical Abstracts Service, describes rules for a name-editing routine that has been developed and programmed.[147] The program performs a character-by-character analysis of a chemical name, and automatically inserts capitals and italics, corrects some punctuation and spacing errors, and expands certain alphabetic abbreviations and chemical line formulas. It is a serially reusable subprogram written for the IBM S/360 in Basic Assembler Language.

Dyson suggests rules for the formation of a general scheme of scientific word lists for computer use, based on analyses of the language used by chemists and biochemists and a classification of their vocabularies.[53] The development of such special vocabularies aims at a basis for index-entry composition and programmed information retrieval.

Conrow developed a computer program for naming organic bridged compounds. The FORTRAN Language program accepts simple structural information on the compounds and uses it to generate the correct names according to the Baeyer system of nomenclature for bridged polycyclics. Conrow was able to name 500 polycyclic systems in about 2 1/2 hours on an IBM 1410. For simplification, the program omits functional and substituent groups; this limits it solely to hydrocarbon skeletons. The basic structural information needed for the program is fed into the computer as a connection matrix. Two or more equivalent names are possible with some symmetrical compounds; the program prints out each possibility with a message noting their symmetry.

Chemical Abstracts Service is developing registration and naming techniques for polymers.[175] Some polymers are being registered on an experimental basis by manual procedures that simulate computer procedures; names, Registry numbers, molecular formulas, and appropriate source indicators are machine recorded but structural information is not. Computer programming for structural information is being undertaken; registration of polymer structures will then be accomplished on a routine basis.

A vital element in the possible listing of stored chemical structure information in various classified arrangements is a recognizable chemical name. Special practices must be adapted for the transcription of such names into the limited typography of punched cards and standard computers. Barnard and associates have studied this transcription problem and have recommended restriction of the characters employed to capital letters, Arabic numbers, and four additional symbols, —, &, /, and *, all of these usually being available on tabulating and electronic processing equipment.[14] These contractions will facilitate the storage of chemical notations.[9]

Chemical Abstracts Service is devising procedures for automatically converting systematic names or organic compounds into atom-bond connection tables which can be manipulated by computer.[199] A dictionary has been prepared of word roots used in the names and in step-by-step procedures for so converting names. The procedures written are applicable to the majority of names of carbon compounds, and the preparation of computer programs has been undertaken. The relationship between structural diagrams and systematic chemical nomenclature provides the link that permits direct computer translation for use in automatic registration. The translation procedures described are applicable to a wide range of systematic nomenclature. When a name is fully processed and converted to a structural record, a molecular formula can be calculated and compared with the molecular formula input by a chemist; any discrepancy will cause the name to be rejected for review. In such editing checks, there is a close parallel to the diagnostic routines for accuracy that are part of the structure registry operations.[114]

The procedures have been tested on two statistically designed samples of CA names.[199] Procedures are applicable to about 60 percent of the names; when studies of names of amino acids, carbohydrates, and compounds containing boron or metals are completed, the procedures should be applicable to 80–85 percent of the names of carbon-containing compounds.

The over-all objective of the CAS project is to develop methods for machine translation of elements of systematic chemical nomenclature to other forms of representation. An early objective is to provide direct conversion of CA index names for compounds into a more detailed machine language, specifically into a connection table. Other objectives include simplification of computer handling of files containing compounds by name, and automatic cross-checking of names against structural diagrams entered into the computer system independently.[59]

Seifer has proposed a set of rules for the mechanical translation of the names of inorganic compounds into the corresponding formulas.[159] The proposed algorithm is applicable to Russian terminology in chemical compounds, but may be applied to English as well.

Dyson proposed a cluster of algorithms relating the nomenclature of organic compounds to their structure matrices and ciphers.[51] A name applied to a complete structure is referred to as a morpheme; parts of names used to modify a morpheme are termed semantemes. Both morphemes and semantemes are deposited in random access storage so that their matrices are available on call. The task to be accomplished is the compilation of a dictionary of morphemes and

semantemes respectively, and programming the systematic interaction of the two groups.

Seifer has described a graphical method of machine language for storing chemical information.[160] The method, based on the construction of composition–property graphs, is used to describe vast amounts of material accumulated in the study of such systems. These graphical operations differ basically from those using linear and structural chemical formulas.

VII

MACHINE DEVELOPMENTS IN CHEMICAL INFORMATION SYSTEMS

*

In a comprehensive system for processing chemical data, there is a real need for input and output of conventional molecular or empirical formulas, various forms of nomenclature, structural-formula diagrams, and considerable generic information and data. The systems analyst was, until recently, limited in his choice of output hardware. This restraint led to the use of numerous and often cumbersome coding techniques and printing conventions that, in many instances, must be learned by the chemists themselves.

Various projects have provided a basis for improved methods of displaying structural diagrams. At Walter Reed Army Institute of Research, a chemical typewriter has been developed that can produce structural diagrams and automatically record alphanumeric and chemical symbols and their cartesian coordinates on punched paper tape. The formatted record on punched paper tape can be manipulated by a computer. At Dow, techniques and conventions have been tested that produce displays of chemical structure diagrams with a high-speed printer. Wiswesser's Dot Plot program also allows display of a structural-formula diagram and notation with a high-speed printer.

The earlier work of Opler and Baird in depicting chemical structure diagrams on a computer-controlled cathode ray tube, as described in Survey of Chemical Notation Systems, gave much promise for the future applicability of such devices. Use of photocomposition to reproduce chemical information and data is of special interest in the publication of the findings of chemists and chemical engineers. However, the techniques developed for these methods of reproduction

are applicable to almost all other branches of information technology where there is a requirement for high-quality multifont and graphic display.

A. INPUT TECHNIQUES AND EQUIPMENT

An early objective in the development of computer-based systems for handling chemical information has been the provision of means for generating and maintaining a file of representations of chemical compounds. Since the college education of chemists has been largely in graphic terms, and the current literature abounds in such structural diagrams, alternative structure descriptions are far more difficult for chemists to learn and use. In these circumstances, the graphic formula often constitutes the primary identification of the individual chemical compound, a situation that has led to the introduction of special tools and facilities into operations for handling the information.

Among such tools are the special typewriters, of varying degrees of sophistication, that allow direct input of structural formulas. An example is the typewriter in use at CIBA Pharmaceutical Company, which has a special key that positions the standard numbers below the line, eliminating the need for frequent rolling of the carriage by hand.[123] At the right end of the same row of keys is a versatile "benzene ring" key, used in typing the general form of rings. At the right end of the third row is a key producing a long vertical line that serves to type the side of a ring in one stroke. With these special keys, little hand work is needed to complete graphic formulas.

At Walter Reed Army Institute of Research, economical input has been achieved by the use of the chemical typewriter developed there. The machine can print up to 128 different characters, can backspace, and can also be equipped to reverse platen-advance. All of these functions are operated under keyboard control. The method by which the input is created also turns out to be a method for obtaining a chemical code; that is, the characters and coordinates generated constitute an atom-by-atom code.[126] Important to the operation of the typewriter is the production of hard copy (sheets or 3 × 5 cards) for the typist to inspect as a guide to making corrections. This feature also assists the person responsible for the regulation of the input to check on errors discovered only after later processing, so that the appropriate corrections can be made. Since some incorrectly drawn structures will require consultation with the submitter of the compound, this production of hard copy at the time of input is exceedingly useful.[96]

On the Mergenthaler-Friden typewriters presently used by the

Army for the input of chemical structures, special chemical characters were provided, not only to facilitate the typing of chemical structures, but also to optimize the speed and convenience of such typing. Chemical structures are not typed linearly, but in two dimensions. In a typical typed chemical structure, the carbon atoms in the ring are represented by dots typed at locations the centers of which correspond to the centers of capital letters C in the completely filled-in structural diagram. If two such dots lying on the same line are to be connected, the connecting bond will be a dash crossing the center of the location of a potential character. But if two dots on successive lines are to be connected, the connecting dash must necessarily be placed between the locations of two potential characters.

Vertical and slanted bonds can be placed between two writing lines according to two methods. On the Army's chemical typewriter, enough space is available on the type slugs to shift a character to a position where it will be typed between two lines.[57] On the Shell-Dura machine this extra space is not available, and the platen is moved instead of the character. One keystroke is used to position the character, and another actually to type it. Although the relative frequency of striking function-keys in the above two methods depends on the material being typed, in practice the second method, where the platen must be moved to interline positions, requires more keystrokes.

The Shell Development Company typewriter will type the basic record of a chemical-structure system and concurrently provide an input tape for computer processing.[135] Such a basic record, for example a 5" x 8" card, may include such items as the registry number, chemical structure, nomenclature, molecular formula, method of preparation, property data, and other information. Symbols for both text and structure are obviously required. A special character set has been designed requiring only nine special symbols, with half-line spacing, to type the structures of most chemical compounds. Additional symbols have been provided for typing stereo configurations. Early plans at Shell do not include the handling of stereo concepts in the computer, but they are provided for graphically, with a view to the future.

The paper tape from the Shell typewriter can be fed to the computer and processed for standardization purposes to a connection table having the form of a triangular matrix. The atoms are placed along the diagonal and bond values are indicated at the intersection of the rows and columns of the atoms joined.

Shell's Chemical Typewriter is distinguished from others by having coded keys for INDEX and BACK INDEX, which respectively move the paper one half-line up or one half-line down without car-

riage return. The options for manual adjustment of the line spacing caused by a carriage return are still available. The typist is instructed to signal the beginning of a structural formula, then copy it in any convenient order, and signal the end of the structure. The rest is done by the computer. The typewriter tape may be used in its original form to retype the structure; no intermediate computer processing is necessary.

Like the symbol set, the INDEX-BACK INDEX feature will work on any typewriter. However, choosing a typewriter with fully interchangeable typing elements makes it unnecessary to commit a typewriter exclusively to chemical information work.

Since January 1966, Shell Research Ltd. has been using the chemical structure typewriter developed at the Emeryville Laboratories.[42,43] At present this equipment is used primarily as an input/output tool for structural formulas in conjunction with the cipher-based system. (The paper-tape image is stored with the compound number.) The next step is to use it for error-detection through matching of the atom matrices generated by the computer from the cipher, on the one hand, and from the typewriter-tape input on the other. Eventually, most of the search queries may be entered through the structure-typewriter, and the response received either through the typewriter, through a printer with special character set, or by means of visual (CRT) display. The Shell-Dura chemical typewriter is also used at Eli Lilly and Company to print structures for dissemination.[141]

Chemical Abstracts Service has utilized two basic methods to enter chemical structures into the Registry System.[3] The first is through the use of connection tables, followed by input with the Mohawk 1101 Data Recorder. In this process, a clerk numbers each nonhydrogen atom in the structure diagram, then keyboards a connection table that lists each atom by number and indicates the atom(s) to which each atom is bonded, along with the type of bond involved. The computer then converts the nonunique, yet unambiguous connection table developed by the clerk into one that not only remains unambiguous, but is also unique. This unique table is stored in the computer.

The second method for registering structures utilizes structure typewriters. Here, a clerk copies the chemical structure directly, using the typewriter, and the computer converts the structure to a unique connection table identical with that obtained by the method described in the preceding paragraph. For procedural reasons and because it should be less expensive, entry of a chemical structure by direct typing is preferable to entry by connection tables. Chemical Abstracts Service utilizes the Invac system, which unites an IBM

Selectric Typewriter with a Mohawk 1181 Data Recorder. This system records the structure directly on magnetic tape. It is expected that some structures will always be registered by connection tables since they are beyond the capacity of the structure typewriter.

Two prototypes of a modified Friden Felxowriter have been built at Smith Kline and French Laboratories (SKF), producing two- and three-dimensional chemical structures based on 38 special characters. Rings produced are of the conventional type and have no gaps. Because of oversized strokes used, only six print strokes are needed to produce a hexagon, facilitating input speed and good graphics. The machines have a reverse-index feature on the keyboard, like the Walter Reed and Shell machines previously described.[72]

Because of the oversized characters in the SKF font, it is not possible to put them on a high-speed chain or drum printer. However, an algorithm has been written to convert SKF font by computer to a dot-bond output, and the latter can be printed as with the Army typewriter. The Flexowriter also has a computer interface permitting operation on line or to a magnetic tape buffer, thus bypassing paper tape if desired.

The good graphics of the SKF system has led to the exploration of a system of composition of structures for printing. The Friden paper tapes are run through a Photon program conversion on the IBM 360/40, and the resulting output tape will drive the Photon 901 to print graphic-quality structures. The Photon matrices have been prepared, and the program is being tested.[72]

Additional experiments are under way to put the SKF font on the Videocom 830 CRT composition device, which will permit output speeds of up to 6,000 characters per second. The purpose of these experiments is to permit full-page composition, including both text and structures.

B. DISPLAY DEVICES FOR COMPUTER-BASED SYSTEMS

The maxim, "one picture is worth a thousand words," is applicable to many complex organic compounds. A report by Levinthal describes model building in which biochemists can study giant molecules displayed on a screen by a computer to gain information about the structure.[116] More generally, chemists often would prefer to receive, as a part of an answer to a search question, a visually recognizable representation of the chemical structures. Several approaches exist for producing a visual display for such purposes.[132] A computer output can be secured corresponding to a previous input of punched tape resulting from the keyboarding of chemical struc-

tures on specially modified electric typewriters. Of greater long-term interest is the mechanical synthesis of a visual display from other information stored in the computer, notably from a line notation or connectivity table. There exist computer programs that convert a Wiswesser Line Notation input to two-dimensional structural formulas for acyclic and cyclic compounds with a conventional line-printer.[90] The program starts from stored connectivity-table records, which in turn are computer-generated from WLN input. The on-line computer generation of visually pleasing displays of chemical structures is an area for active study and research.

In an attempt to solve the display problem, the General Electric Company experimented, for the CIDS program, with the SC-4020 high-speed CRT to determine the minimum feasible size of characters that would allow economically for adequate font resolution.[30] It was determined initially that character sets for Arabic numerals, uppercase and lowercase Greek and Latin letters, and extensive sets of chemical and special symbols would be desired. In addition, maximum flexibility was required in alphanumeric superscripting and subscripting in the text displayed.

A digital computer was used to create instructions off-line to the CRT for display of formatted system-output records. In this way, machine-editing functions could be changed, or new editing functions could be added to the computer programs. The flexibility afforded by this technique can eliminate the need for modification or construction of special display equipment.

The major disadvantage of the CRT is that unless it is coupled with Xerographic or photographic equipment, or to suitable typewriting equipment, the copy is "soft." The only solutions are to attach such equipment, which would increase the cost of the console, or to rely upon another output device for making hard copy of the structures. The least expensive would be an ordinary electric typewriter with an appropriate set of characters. The standard two-shift teletypewriter would serve the purpose.

While excellent structure display can be obtained by using a chemical typewriter for computer input and a line-printer for output, chemical typewriter input to a system is relatively slow and results in a long record. Work at Canadian Industries Ltd. was designed to establish the feasibility of using the compact record derived from the Wiswesser Line Notation to generate an acceptable structure display for output on the line-printer.[188] An advantage of this approach is that one record serves the dual purpose of search and display. A computer program to generate chemical structure diagrams from the Wiswesser notation has been written and successfully tested, as mentioned previously in this report. A two-dimen-

sional picture can be generated, using the compacted-matrix form of the notation.[90]

However, a program for generating structure display must compete economically with the alternative method of storing a separate display record on tape. At some point, the computer generation will be more expensive than creating and storing a separate record, and this limitation must inevitably lead to a compromise. In a number of cases, a certain amount of difficult draftmanship is required to generate a structure in an acceptable form. It is proposed that a separate display record would then be created, since a display generated from the connectivity matrix would be impractical.

The steps involved in converting the compacted matrix to a printed structure are as follows: (1) deriving the connection paths for the Wiswesser chemical units, (2) translating the Wiswesser chemical units into their normal atomic representation, and (3) plotting each atom and bond according to a free plotting routine that takes into consideration (a) the direction changes required at branching points, (b) the direction changes required to plot points on a ring, and (c) the ability to modify a particular tracking route when an overwriting conflict is likely to occur.[90]

The first two steps deal with rearrangement and translation of Wiswesser symbols into a format suitable for plotting. Translation of Wiswesser symbols into conventional atom representation requires a look into one of the other two conversion tables—one for forward plotting and the other for reverse plotting. The coordinates of points required to print a structure are derived and plotted in a grid area defined in the memory of the computer. Branching atoms, such as N, P, and secondary or tertiary C, represent points from which directional changes are required. In plotting from such a point the computer has available eight possible tracking routes, and selects them in a clockwise order. Having established the direction of approach and the desired angles required between branches, the tracking routes are obtained from a table that provides all possible directional changes from a praticular track.

The routine that plots points on a ring considers the center of the ring to be a branching point with the same eight possible tracking routes. Depending on the size of the ring and direction of approach, horizontal or vertical rings are plotted by a selection of the available direction. In plotting a fused-ring system, the coordinates of the lowest ring-fusion atom in the previous ring are noted and are used to calculate the ring center or point of origin of the next ring.

An additional routine notes the lowest and highest coordinates plotted on each line. These data are required only at a branching point or ring center since a direction change might result in over-

writing. If a particular path would lead to this overlap problem, the bond leading to the branching point or ring is extended or stretched until the problem area is cleared.

Experience gained so far indicates that the concept of structure generation from notation is a feasible one.[90] In many cases, the draftmanship for acceptable representation of complex molecules would be impractical to generate from any linear record. For such structures it is intended to create and hold a separate display file derived from a chemical typewriter.

VIII

CONCLUDING REMARKS

*

This review is based on the published and unpublished literature on chemical structure information handling that appeared during the period from 1962 through late 1968. Some aspects of the subject matter were documented more extensively than others in this period, although they were not necessarily more significant components in the total picture of efforts in the field. The literature review was augmented by the personal knowledge and experiences of the members of the Committee on Chemical Information of the National Research Council.

The source material was catagorized and arranged to summarize developments and accomplishments and to present a comprehensive and unified view of the subject at this point in time. The rapid increases in capability, versatility, and availability of electronic data-processing equipment, and the increasingly prominent role of such equipment in systems design have introduced new dimensions to research and development efforts in this area. These changes will affect future trends in a manner difficult to predict at this time.

REFERENCES AND BIBLIOGRAPHY

Bibliographic entries are designated by an asterisk.

*

1. Aims, A. Survey of information needs of physicists and chemists; the report of a survey undertaken in 1963-4, in association with Professor B. H. Flowers, on behalf of the Advisory Council on Scientific Policy. J. Doc., 21(2):83–112.
2. Anderson, R., E. Marden, and B. Marron. File organization for a large chemical information system. U.S. Department of Commerce, National Bureau of Standards. Institute for Applied Technology. Technical Note 285. 1966, 17 p.
3. Annual report to the National Science Foundation on contract NSF-C414 Task I, 1 June 1966 through 31 May 1967. American Chemical Society. Chemical Abstracts Service. Annual report, Columbus, Ohio. 1967, 46 p. + appendix.
4. Armitage, J. E., and M. F. Lynch. Automatic detection of structural similarities among chemical compounds. J. Chem. Soc., Section C: Organic Chemistry, 521–528. 1967.
5. Armitage, J. E., and others. Documentation of chemical reactions by computer analysis of structural changes. J. Chem. Doc., 7(4):209–215. 1967.
6. Baker, D. B. Chemical literature expands. Chem. Eng. News, 44(23):84–87. 1966.
7. Baker, D. B. Chemical Abstracts Service. Chem. Eng. News, 46(16):9A–10A. 1968.
8. Ballard, D. L., and F. Neeland. A computer technique for the retrieval of related chemical structures utilizing a special topological cipher. J. Chem. Doc., 3(4):196–201. 1963.
9. Barnard, A. J., C. T. Kleppinger, and W. J. Wiswesser. Computer-oriented chemical names. J. Chem. Doc., 6(1):48–57. 1966.

10. Barnard, A. J., C. T. Kleppinger, and W. J. Wiswesser. Retrieval of organic structures from small-to-medium sized collections. J. Chem. Doc., 6(1):41–48. 1966.
11. Barnard, A. J., and W. J. Wiswesser. Some innovations in chemical information management. Information Retrieval Letter, 2(6):1–3. 1966.
12. Barnard, A. J., and W. J. Wiswesser. Use of the BATCH number with hand-manipulated files. J. Chem. Doc., 6(3):188–189. 1966.
13. Barnard, A. J., and W. J. Wiswesser. Computer-serviced management of chemical structure information. Lab. Manage., 5(10):34–36, 40–42, 44. 1967.
14. Barnard, A. J., and others. Some techniques for the machine management of small chemical data systems. p. 85–101 In J. P. Mitchell, ed. Proceedings of the Wiswesser Line Notation meeting of the Army chemical information and data systems program, 6–7 October, 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. EASP 400-8 (AD 665 397) Edgewood Arsenal, Maryland. 1968, 212 p.
15. Barnes, R. F. Language problems posed by heavily structured data. Paper presented at an Open Technical Meeting on Design, Implementation and Application of IR-Oriented Languages. Association for Computing Machinery, Computer Language Committee on Information Retrieval. Princeton, N.J. 1961.
16. Bernays, P. M. Statistical data on chemical compounds; final report. U.S. Army. Edgewood Arsenal. Chemical Research and Development Laboratories. (AD 615 488) Edgewood Arsenal, Maryland (1965), 19 p.
17. Bernays, P. M., K. L. Coe, and J. L. Wood. A computer-based source inventory of Chemical Abstracts. J. Chem. Doc., 5(4):242–249. 1965.
18. Bobka, M. E., and J. B. Subramaniam. A computer oriented scheme for coding chemicals in the field of biomedicine. Case Western Reserve University. School of Library Science. Center for Documentation and Communication Research. Comparative Systems Laboratory. Technical Report 11, Cleveland, Ohio. 1967, 22 p.
19. Bonnett, H. T. Chemical notations: a brief review. J. Chem. Doc., 3(4): 235–242. 1963.
20. Bonnett, H. T. The current status of mechanical manipulation of chemical structural information. Paper presented at 151st meeting, American Chemical Society, Division of Chemical Literature. Chicago, Illinois. 1964.
21. Bonnett, H. T. Use of the Wiswesser line notation at the Searle Laboratories: motivation and status. p. 15–23 In J. P. Mitchell, ed. Proceedings of the Wiswesser line notation meeting of the Army chemical information and data systems program, 6–7 October, 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. EASP 400-8 (AD 665 397) Edgewood Arsenal, Maryland. 1968. 212 p.
22. Bouman, H. Computer program for the LINCO system. J. Chem. Doc., 5(1):14–24. 1965.
23. Bowman, C. M., F. A. Landee, and M. H. Reslock. A chemically oriented information storage and retrieval system (I): storage and verification of structural information. J. Chem. Doc., 7(1):43–47. 1967.
24. Bowman, C. M., and others. Automatic generation of structural fragment codes from the Wiswesser Line Notation for rapid structure searches.

p. 49–56 In J. P. Mitchell, ed. Proceedings of the Wiswesser Line Notation meeting of the Army chemical information and data systems program, 6–7 October, 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. EASP 400-8, (AD 665 397) Edgewood Arsenal, Maryland. 1968, 212 p.

25. Bowman, C. M., and others. A chemically oriented information storage and retrieval system (II): computer generation of the Wiswesser notations of complex polycyclic structures. J. Chem. Doc., 8(3):133–138. 1968.

*26. Brasie, W. C., and D. W. Liou. Computer estimation of physical properties using structure coded compound inputs. Paper presented at fifty-seventh annual meeting, American Institute of Chemical Engineers. Symposium on systems for estimation of physical properties. Boston, Massachusetts. 1964.

27. Brasie, W. C., and D. W. Liou. Estimating physical properties: chemical structure coding. Chem. Eng. Progr. 61(5):102–108. 1965.

28. Brunner, R. G., and others. Chemical data compilation analysis survey. Science Communication, Incorporated. Final Report (NSFC-478), Washington, D.C. 1967, 29 p. + appendices A-F.

29. Buchanan, B., and G. Sutherland. Heuristic Dendral: a program for generating explanatory hypotheses in organic chemistry. Stanford University. (AD 673 382) Stanford Artificial Intelligence Project. Memo AI-62, Stanford, California. 1968, 79 p.

30. Burger, J. B. High speed display of chemical nomenclature, molecular formula and structural diagram. (AD 460 820) U.S. Army Missile Command. Redstone Scientific Information Center. Report CIDS-4, Redstone Arsenal, Alabama. 1964, 19 p.

*31. Burger, J. B., and W. J. Wilson. A review of selected methods of machine manipulation of chemical structures. U.S. Army Missile Command. (AD 451 700) Redstone Scientific Information Center. Report RSIC-287, Redstone Arsenal, Alabama. 1964, 58 p.

32. Cahn, R. S. An introduction to the sequence rule; a system for the specification of absolute configuration. J. Chem. Ed. 41(3):116–125. 1964.

33. Cockayne, A. H., and E. Hyde. Prime number coding for information retrieval. Comp. J. 3(1):21–22. 1960.

34. Coding chemical compounds. (News item) Pharm. J. 196(5356):686. 1966.

*35. Coe, K. L., and others. Procedures for assessing errors. J. Chem. Doc. 6(4):129–132. 1966.

*36. Computer aids search for R & D chemicals. Chem. Eng. News 46(38):26–27. 1968.

*37. Computer program names bridged organics: Fortran language system generates Baeyer names from simple structural information. Chem. Eng. News 44(51):84–86. 1966.

38. Conference on Mechanized Processing of Chemical Information. March 5–8, 1964. Airlie House, Warrenton, Virginia. Washington, D.C.: National Academy of Sciences–National Research Council, 1964. National Science Foundation Contract NSF-C310. (Summary report prepared by Dr. W. H. Waldo, approved by CMMHCI, Oct. 3, 1964.)

*39. Conrow, K. Computer generation of Baeyer system names of saturated bridged bicyclic, tricyclic and tetracyclic hydrocarbons. J. Chem. Doc. 6:206–213. 1966.

40. Cossum, W. E., M. E. Hardenbrook, and R. N. Wolfe. Computer generation of atom-bond connection tables from hand-drawn chemical structures. American Documentation Institute. Proceedings, 27th annual meeting, Philadelphia, Pennsylvania, October 5–8, 1964. vol 1:269–275. 1964.
41. Cossum, W. E., M. L. Krakiwsky, and M. F. Lynch. Advances in automatic chemical substructure searching techniques. J. Chem. Doc. 5(1):33–35. 1965.
42. Dammers, H. F. Computer-based chemical information system. New Sci. 31:325–327. 1966.
43. Dammers, H. F. Computer handling of literature information and research data in an industrial research establishment. (Paper presented at 36th International Congress on Industrial Chemistry. Brussels, 1966.)
44. Dammers, H. F., and D. J. Polton. Use of the IUPAC notation in computer processing of information on chemical structures. J. Chem. Doc. 8(3): 150–160. 1968.
45. Davenport, W. C. The Chemical Abstracts Service computer based chemical information processing system. Ohio State University. Chemical Abstracts Service. Columbus, Ohio. 1965. 10 p.
46. Davenport, W. C. CAS computer-based information services. Datamation p. 33-38. March 1968.
47. De Laet, F. C. R. Coding and recording of chemical flow-sheets for mechanical search. Rev. Internat. Doc. 32(3):91–98. 1965.
48. De Maine, P. A. D., and B. A. Marron. The solid system (I): a method for organizing and searching files. In Information Retrieval: A Critical View, Washington, D.C. Thompson Book Company. 1967. p. 243–282.
49. de Mott, A. N. Interpretation of organic chemical formulas by computer. In American Federation of Information Processing Societies. AFIPS Conference Proceedings. (1968 Spring Joint Computer Conference, Atlantic City.) Washington, D.C. Thompson Book Company. 1968. Vol. 32 p. 61–65.
50. Dyson, G. M., and others. Mechanical manipulation of chemical structure: molform computation and substructure searching of organic structures by the use of cipher-directed, extended and random matrices. Inform. Storage and Retrieval. 1(2-3):66–69. 1963.
51. Dyson, G. M. A cluster of algorithms relating the nomenclature of organic compounds to their structure matrices and ciphers. Inform. Storage and Retrieval. 2(3):159–199. 1964.
52. Dyson, G. M. Generic (or Markush) groups in notation and search programs, with particular reference to patents. Inform. Storage and Retrieval. 2(2):59–71. 1964.
53. Dyson, G. M. Computer input and the semantic organization of scientific terms (I). Inform. Storage and Retrieval. 3:35–115. 1967.
54. Dyson, G. M. Modifications and abbreviations recommended for computer and visual handling of the IUPAC notation. (Communication from the author)
55. Eisman, S. H. A Polish-type notation for chemical structures. J. Chem. Doc. 4(3):186–190. 1964.
56. Feeman, J. J. A novel organizational code for organic structures based on functional groups. J. Chem. Doc. 6(3):184–187. 1966.

57. Feldman, A. A chemical teletype. (Paper presented at 154th meeting, American Chemical Society. Chicago, 1967.)
58. Feldman, A. A proposed improvement in the printing of chemical structures, which results in their complete computer codes. Amer. Doc. 15(3): 205–209. 1964.
59. Final report to the National Science Foundation on research and development supported by Grant NSF-GN 382, January–May 1965. (PB 168 500) American Chemical Society. Chemical Abstracts Service. Columbus, Ohio. 1965.
*60. Fraction, G. F. Substructure searching on notations. p. 69–79 In J. P. Mitchell, ed. Proceedings of the Wiswesser line notation meeting of the Army chemical information and data systems program, 6–7 October, 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. (AD 665 397) Edgewood Arsenal Special Publication EASP 400-8, Edgewood Arsenal, Maryland. 1968, 212 p.
61. Fraction, G. F., J. C. Walker, and S. J. Tauber. Connection table from Wiswesser Line Notation: a partial algorithm. p. 139–195 In J. P. Mitchell, ed. Proceedings of the Wiswesser Line Notation meeting of the Army chemical information and data systems program, 6–7 October 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. (AD 665 397) Edgewood Arsenal Special Publication EASP 400-8, Edgewood Arsenal, Maryland. 1968, 212 p.
62. Fraction, G. F., J. C. Walker, and S. J. Tauber. Connection tables from Wiswesser chemical structure notations—a partial algorithm. U.S. Department of Commerce. National Bureau of Standards. Technical Notes 432, Washington, D.C. 1968, 25 p.
*63. Frome, J. Searching chemical structures. J. Chem. Doc. 4(1):43–45. 1964.
64. Fugmann, R. Experiences with a faceted classification in organic chemistry using computers. p. 341–363 In P. Atherton, ed. Classification Research; Proceedings of the 2nd International Study Conference. Elsinore, Denmark. 1965.
*65. Fugmann, R., U. Dölling, and H. Nickelsen. A topological approach to the problem of ring structures. Angewandte Chemie. International Edition in English. (Weinheim) 6(9):723–818. 1967.
66. Garfield, E. Generic searching by use of rotated formula indexes. J. Chem. Doc. 3:97–103. 1963.
67. Gelberg, A. Rapid structure searches via permuted chemical line notations (IV): a reactant index. J. Chem. Doc. 6(1):60–61. 1966.
*68. Gelberg, A. Permutations and classification numbers. p. 80–84 In J. P. Mitchell, ed. Proceedings of the Wiswesser Line Notation meeting of the Army chemical information and data systems program, 6–7 October, 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. Edgewood Arsenal Special Publication EASP 400-8, (AD 665 397) Edgewood Arsenal, Maryland. 1968, 212 p.
*69. Gelberg, A. Quick scan and symbols. p. 43–48 In J. P. Mitchell, ed. Proceedings of the Wiswesser Line Notation meeting of the Army chemical information and data systems program, 6–7 October, 1967. U.S.

Army Edgewood Arsenal. Technical Support Directorate. Edgewood Arsenal Special Publication EASP 400-8, (AD 665 397) Edgewood Arsenal, Maryland. 1968, 212 p.

70. Gibson, G. W. The Wiswesser line-notation: an introduction. U.S. Army. Edgewood Arsenal. Chemical Research and Development Laboratories. Technical memorandum CRDL 7-3, (AD 624 525) Edgewood Arsenal, Maryland. 1965, 31 p.
71. Gluck, D. J. A chemical structure storage and search system developed at DuPont. J. Chem. Doc. 5(1):43–51. 1965.
72. Gordon, M. The potential impact of chemical typewriters on documentation. (Translated from: Die Pharmazeutische Industrie. 28:893–897. 1966) Philadelphia, Smith Kline French Laboratories, [1967?] 11 p.
73. Gould, D., E. B. Gasser, and J. F. Rian. ChemSEARCH—an operating computer system for retrieving chemicals selected for equal, analogous or related character. J. Chem. Doc. 5(1):24–32. 1965.
74. Grady, F., and E. Hyde. Mechanized information retrieval at ICI pharmaceuticals division. Mfg. Chem. Aerosol News. October:1–5. 1964.
75. Granito, C. E., and others. Rapid structure searches via permuted chemical line notations (II): A key-punch procedure for the generation of an index for a small file. J. Chem. Doc. 5(1):52–55. 1965.
76. Granito, C. E., and others. Rapid structure searches via permuted chemical line notations (III): A computer-produced index. J. Chem. Doc. 5(4): 229–233. 1965.
77. Granito, C. E., D. E. Renard, and L. A. Holly. Use of Wiswesser Line Notation for determining duplicate chemical structures. J. Chem. Doc. 6:252–253. 1966.
78. Granito, C. E. Use of the Wiswesser Line Notation for registering compounds. p. 35–37 In J. P. Mitchell, ed. Proceedings of the Wiswesser Line Notation meeting of the Army chemical information and data systems program, 6-7 October, 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. Edgewood Arsenal Special Publication EASP 400-8, (AD 665 397) Edgewood Arsenal, Maryland. 1968, 212 p.
79. Haefele, C. R., and J. F. Tinker. Some unusual features of a chemical retrieval system used in the Eastman Kodak Company. J. Chem. Doc. 4(2):112–115. 1964.
80. Hayward, H. W. A second look at chemical notation systems in view of projected machine interconversion of cipher forms. p. 59–60 In Automation and Scientific Communication; short papers of the 26th Annual Meeting of the American Documentation Institute, Chicago, 1963. Washington, American Documentation Institute, 1963. Part 1.
81. Hayward, H. W., and others. Some experience with the Hayward linear notation system. J. Chem. Doc. 5(3):183–189. 1965.
82. Hayward, H. W., and S. J. Tauber. The HAYSTAQ experiment. (preliminary version, unpublished) U.S. Department of Commerce. Patent Office. Washington, 1965.
83. Hiz, H. A linearization of chemical graphs. J. Chem. Doc. 4(3):173–180. 1964.
84. Hoffman, W. S. An integrated chemical structure storage and search system operating at DuPont. J. Chem. Doc. 8(1):3–13. 1968.

85. Horner, J. K. Low-cost storage and retrieval of organic structures by permuted line notations: small collections. p. 25–33 In J. P. Mitchell, ed. Proceedings of the Wiswesser Army chemical information and data systems program, 6–7 October, 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. Edgewood Arsenal Special Publication EASP 400-8, (AD 665 397) Edgewood Arsenal, Maryland, 1968, 212 p.
86. Horowitz, P., and E. M. Crane. HECSAGON: A system for computer storage and retrieval of chemical structure. Eastman Kodak Company, Rochester, New York. 1961, 33 p.
87. Huber, M. L. Chemical structure codes in perspective. J. Chem. Doc. 5(1):4–8. 1965.
88. Hyde, E., and others. Conversion of Wiswesser notation to a connectivity matrix for organic compounds. (Paper presented at 153rd National Meeting, American Chemical Society, Division of Chemical Literature. Miami Beach, 1967.)
*89. Hyde, E. Computer generated open ended fragment code. p. 57–67 In J. P. Mitchell, ed. Proceedings of the Wiswesser Line Notation meeting of the Army chemical information and data systems program, 6–7 October, 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. Edgewood Arsenal Special Publication EASP 400-8, (AD 665 397) Edgewood Arsenal, Maryland. 1968, 212 p.
90. Hyde, E., and L. Thomson. Structure display. J. Chem. Doc. 8(3):138–146. 1968.
91. Ihndris, R. W. Chemical structure fragmentation for use in a coordinate index retrieval system. J. Chem. Doc. 4(4):274–278. 1964.
92. Index Chemicus Registry System; a sample of chemical compounds reported in the December 1967 INDEX CHEMICUS . . . encoded in the Wiswesser Line Notation. Philadelphia, Institute for Scientific Information, [1968?] 16 p.
93. Indexing changes slated for CA's Vol. 66. Chem. Eng. News 46(31):47. 1968.
94. Information services from Chemical Abstracts Service, 1969. (brochure) Columbus, Ohio, American Chemical Society. Chemical Abstracts Service. [1969], 32 p.
95. Jacobus, D. P., and D. E. Davidson. Processing large numbers of chemical compounds and biological data. Drug Information Association. Proceedings. Vol 2:199–211. 1966.
96. Jacobus, D. P., and others. Experience with the mechanized chemical and biological information retrieval system. (Paper presented at 154th Meeting, American Chemical Society, Division of Chemical Literature. Chicago, 1967.)
97. J. T. Baker BATCH directory: a classified index of compounds based on structural and atomic elements. Phillipsburg, New Jersey. 1965. 39 p.
98. Kirschner, S., S. H. Kravitz, and J. Mack. The application of computers to the retrieval of selective information regarding the anticancer activity of coordination compounds. J. Chem. Doc. 6(4):213–217. 1966.
99. Kokoropoulos, P., and F. Scheffler. The "basic structure-substituent connector" concept for handling organic compounds for information retrieval purposes. American Documentation Institute. Proceedings, twenty-seventh annual meeting, Philadelphia, 1964. Vol. 1:399–402. 1964.

100. Kokoropoulos, P., and S. V. Kanakkanatt. An indexing system and code for polymers. J. Chem. Doc. 8(3):179–187. 1968.
101. Koller, H. R. The chemists' approach to the information problem, or how and why chemists are using information machines and why they are not. (Paper presented at Western Electronics Show and Convention, Los Angeles, 1966.)
102. Krakiwsky, M. L., W. C. Davenport, and R. W. White. Searching for subsets in machine records of chemical structures at Chemical Abstracts Service. (Paper presented at the 1965 Congress of the International Federation for Documentation, Washington, 1965.)
103. Kulpinski, S., and others. A study and implementation of mechanical translation from Wiswesser Line Notation to connection table, Vol. I. University of Pennsylvania. Engineering Schools, Engineering Research. Annual report, Contract NSFC-467 (PB 177 292) Philadelphia, 1967, 74 p. + appendices A–D.
104. Kulpinski, S., and others. A study and implementation of mechanical translation from Wiswesser Line Notation to connection table, Vol. II: Program macro flow charts. University of Pennsylvania. Engineering Schools-Engineering Research, Annual report, Contract NSF-C-467 (PB 177 292) Philadelphia, 1967. 43 p.
105. Landee, F. A. Computer methods of handling files of chemically oriented information. Dow Chemical Company, Midland, Michigan. 1965.
106. Lederberg, J. DENDRAL-64, A system for computer construction, enumeration and notation of organic molecules as tree structures and cyclic graphs. Part I: Notational algorithm for tree structures. (Interim report to National Aeronautics and Space Administration. NASA CR-5729. 1964.)
*107. Lederberg, J. Topological mapping of organic molecules. Proc. Nat. Acad. Sci., U.S. 53(1):134–139. 1965.
108. Lee, R. W. H., and others. TOPKAT (Topological Kind of Attack); a computer based substructure search on generic structures. U.S. Department of Commerce. National Bureau of Standards. Report, Contract NSF-AG-69, Washington, 1968. 39 p. + appendices I and II, and Figures 1–18.
109. Lefkovitz, D., and C. T. Van Meter. An experimental real time chemical information system. J. Chem. Doc. 6(3):173–183. 1966.
110. Lefkovitz, D. The impact of third generation ADP equipment on alternative chemical structure information systems. (Paper presented at 153rd National Meeting, American Chemical Society. Miami Beach, 1967.)
111. Lefkovitz, D., and R. V. Powers. A list-structured chemical information retrieval system. In Information Retrieval: a critical view. Thompson Book Company, Washington. 1967.
112. Lefkovitz, D. Substructure search in the MCC system. J. Chem. Doc. 8(3):166–173. 1968.
113. Leiter, D. P., H. L. Morgan, and R. E. Stobaugh. Installation and operation of a registry for chemical compounds. J. Chem. Doc. 5(4):238–242. 1965.

114. Leiter, D. P., and H. L. Morgan. Quality control and auditing procedure in the Chemical Abstracts Service compound registry. J. Chem. Doc. 6(4):226–229. 1966.
115. Leiter, D. P., and L. H. Leighner. A statistical analysis of the structure registry at Chemical Abstracts Service. (Paper presented at 154th National Meeting, American Chemical Society. Chicago, 1967.)
116. Levinthal, C. Molecular model-building by computer. Sci. Amer. 214(6): 42–52. 1966.
117. Lissant, K. J. A unified method of delineating polymeric species. J. Chem. Doc. 3:103–113. 1963.
118. Lynch, M. F. Subject indexes and automatic document retrieval; the structure of entries in Chemical Abstracts subject indexes. J. Doc. 22(3):167–185. 1966.
119. Lynch, M. F. Conversion of connection table descriptions of chemical compounds into a form of Wiswesser notation. J. Chem. Doc. 8(3):130–133. 1968.
120. Lynch, M. F. Storage and retrieval of information on chemical structures by computer. Endeavor 27:68–73. 1968.
121. Marden, E. C. HAYSTAQ, a mechanized system for searching chemical information. U.S. Department of Commerce, National Bureau of Standards. Institute for Applied Technology. Technical Note 264, Washington. 1965, 59 p.
122. Marron, B. A., G. R. Bolotsky, and S. J. Tauber. Chemical substructure searching with linear notations. J. Chem. Doc. 6(2):92–95. 1966.
123. Marsh, J. L. Documentation of preclinical research information. Drug Inform. Bull. April/June:86–90. 1967.
124. Mathers, B. L. System Development plan for a national chemical information system. (AD 650 900) Information Management, Incorporated, Washington, 1967, 30 p.
125. Mathers, B. L., and others. System performance specification for a national chemical information system. (AD 650 901) Information Management, Incorporated, Washington. 1967, 124 p.
126. McDonnell, P. M., and R. F. Pasternack. A line-formula notation system for coordination compounds. J. Chem. Doc. 5(1):56–60. 1965.
127. Meadows, E. L. The A.I.Ch.E. physical properties estimation system. (Paper presented at fifty-seventh annual meeting, American Institute of Chemical Engineers. Symposium on systems for estimation of physical properties. Boston, 1964.)
128. Meadows, E. L. Estimating physical properties: The A.I.Ch.E. system. Chem. Eng. Progr. 61(5):93–95. 1965.
129. Method for handling confidential data in the CAS chemical compound registry system. American Chemical Society. Chemical Abstracts Service. Columbus, Ohio. 1966, 46 p.
130. Meyer, E. Mechanization of chemical documentation. Angew. Chem. International Edition in English (Weinheim). 4(4):347–352. 1965.
*131. Meyer, E. Eine topologische Kurzdarstellung chemischer Strukturformeln für die Dokumentation mit elektronischen Rechenanlagen. [a topological coding of chemical structure formulas for documentation with

computer processing capability] Inform. Storage and Retrieval. 2(4): 205–215. 1965.

132. Mitchell, J. P., ed. Proceedings of the Wiswesser Line Notation meeting of the Army chemical information and data systems program, 6–7 October, 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. Edgewood Arsenal Special Publication EASP 400-8 (AD 665 397) Edgewood Arsenal, Maryland. 1968, 212 p.
133. Morgan, H. L. The generation of a unique machine description for chemical structures—a technique developed at Chemical Abstracts Service. J. Chem. Doc. 5(2):107–113. 1965.
134. Morphis, B. B., and others. Computer interpretation of biological and chemical data. J. Chem. Doc. 6(2):77–81. 1966.
135. Mullen, J. M. Atom-by-atom typewriter input for computerized storage and retrieval of chemical structures. J. Chem. Doc. 7(2):88–93. 1966.
136. Munz, J. On the recognition of chemical rings. University of Pennsylvania, Analysis of Chemical Notations Project. Vol. I, 1965. 25 p.
137. Munz, J. The formal analysis of notation systems. p. 197–202 In J. P. Mitchell, ed. Proceedings of the Wiswesser Line Notation meeting of the Army chemical information and data systems program, 6–7 October, 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. Edgewood Arsenal Special Publication EASP 400-8 (AD 665 397) Edgewood Arsenal, Maryland. 1968. 212 p.
138. Munz, J. The formal evaluation of a simple chemical cipher system. University of Pennsylvania. Department of Linguistics. Analysis of Chemical Notations Project. Paper No. 3. 1968, 42 p.
139. Munz, J. Linguistics and information concerning chemical structure. (Communication from the author)
140. Norris, R. C. Route selection for the estimation of properties of pure compounds. (Paper presented at fifty-seventh annual meeting, American Institute of Chemical Engineers. Symposium on systems for estimation of physical properties. Boston, 1964.)
141. Ofer, K. D., and others. A pilot study for the input to a chemical-structure retrieval system. (Paper presented at American Chemical Society Meeting, Division of Chemical Literature, Pittsburgh, 1966.)
142. Ofer, K. D. A computer program to index or search linear notations. J. Chem. Doc. 8(3):128–129. 1968.
143. Olejar, P. D. The interagency chemical information program. (Paper presented at OST Interagency Meeting, March 24, 1966.)
144. Opler, A. A brief survey of topological representations. (Paper presented at 27th Annual Meeting, American Documentation Institute, Philadelphia, 1964.)
145. An overview of worldwide chemical information facilities and resources. National Science Foundation. [Report] for the Joint Study on the Communication of Scientific Information and on the Feasibility of a Worldwide Science Information System of the International Council of Scientific Unions and the United Nations Educational, Scientific and Cultural Organization, (PB 176 160) National Science Foundation, Washington, 1967, 85 p. + appendices A–M.

146. Papers on critical reviews (presented at 155th Meeting, American Chemical Society, Division of Chemical Literature, Symposium on Critical Reviews, San Francisco. 1968). J. Chem. Doc. 8(4):231–245. 1968.
147. Park, M. K. Automatic editing of chemical nomenclature. (Paper presented at 154th National Meeting, American Chemical Society. Chicago, 1967.)
148. Park, M. K., R. E. Stobaugh, and R. J. Zalac. The use of a computer-generated notation for handling compounds in a mechanized system. Ohio State University Chemical Abstracts Service. Columbus, Ohio. 1968, 14 p.
149. Pasternack, R. F., and P. M. McDonnell. Designation of ligand positions in coordination complexes. Inorg. Chem. 4(4):600–602. 1965.
150. Penny, H. A connectivity code for use in describing chemical structures. J. Chem. Doc. 5(2):113–117. 1965.
*151. Petrarca, A. E., M. F. Lynch, and J. E. Rush. A method for generating unique computer structural representations of stereoisomers. J. Chem. Doc. 7(3):154–165. 1967.
152. Pielmeier, G. R. A review of federal chemical information and data systems. (Paper presented at 153rd meeting, American Chemical Society, Miami Beach, 1967.)
153. Renard, D. E. Updating program for the industry liaison office. p. 117–119 In J. P. Mitchell, ed. Proceedings of the Wiswesser Line Notation meeting of the Army chemical information and data systems program, 6–7 October, 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. Edgewood Arsenal Special Publication EASP 400-8 (AD 665 397) Edgewood Arsenal, Maryland. 1968, 212 p.
154. A review of the Chemical Information Program, Part II, July 1, 1966, to June 30, 1967. National Science Foundation. Office of Science Information Service. Chemical Information Unit, Washington. 1967, 35 p.
155. Rule, D. F. The application of the 2280 film recorder to CAS processes. (Report on the eighth Chemical Abstracts Service open forum, held in conjunction with the 154th National Meeting of the American Chemical Society, Chicago, 1967.) American Chemical Society. Chemical Abstracts Service. Columbus, Ohio. 1968, 10 p.
156. Salton, G., and E. H. Sussenguth. Some flexible information retrieval systems using structure matching procedures. p. 587–597 In American Federation of Information Processing Societies. AFIPS Conference Proceedings (1964 Spring Joint Computer Conference) Vol. 25: Baltimore, Spartan Books, 1964.
157. Seidel, M. R. Biodirectional reference array internally derived. (BRAID) Datatrol Corporation. Information Storage and Retrieval Division. Technical Report 1R-6. 1963, 15 p.
158. Seidel, M. P. Chemical BRAID, a technique for totally-indexed structure storage. Datatrol Corporation. Technical Report IR-16, no date, 14 p.
159. Seifer, A. L. Algorithms for converting the names of inorganic compounds into formulae. Inform. Storage and Retrieval. 1:29–40. 1963.

160. Seifer, A. L., G. I. Kleinermann, and V. S. Stein. The principles of construction of a machine language for physicochemical analysis. Inform. Storage and Retrieval. 1(1):13–18. 1963.

*161. Sher, I. H., J. O'Conner, and E. Garfield. Rotadex—A new Index for Generic Searching of Chemical Compounds. J. Chem. Doc. 4(1):49–53. 1964.

162. Silk, J. A. A notation-based fragment code for chemical patents. J. Chem. Doc. 8(3):161–165. 1968.

163. Silverton, E., and R. F. Pasternack. A line formula notation system for coordination compounds (III): Deviations from idealized configurations. U.S. Department of Commerce. National Bureau of Standards. J. Res. NBS. A, Physics and Chemistry. 70A(1):23–27. 1966.

164. Skolnik, H., and A. Clow. A notation system for indexing pesticides. J. Chem. Doc. 4(4):221–227. 1964.

165. Smith, C. User requirements for chemical information and data system (CIDS). U.S. Army. Frankford Arsenal. Research and Development Directorate. Report R-1755 (AD 616 889) Frankford Arsenal, Philadelphia. 1965, 65 p.

166. Smith, E. G. Recent developments with the Wiswesser notation. (Report prepared for the subcommittee on chemical notation systems development and interconversion, Committee on Modern Methods of Handling Chemical Information, National Academy of Sciences–National Research Council. Revised, 1965.)

167. Smith, E. G. The Wiswesser Line-Formula Chemical Notation. McGraw-Hill, New York. 1968.

168. Smith, E. G. Wiswesser notation tutorial. (Paper presented at 155th Meeting, American Chemical Society, Division of Chemical Literature. San Francisco, 1968.)

169. Sneed, H. M. S., J. H. Turnipseed, and R. A. Turpin. A line-formula notation system for Markush structures. J. Chem. Doc. 8(3):173–178. 1968.

170. Sorter, P. F., and others. Rapid structure searches via permuted chemical line-notations. J. Chem. Doc. 4(1):56–60. 1964.

171. Sorter, F. File maintenance and updating procedure. p. 39–42 In J. P. Mitchell, ed. Proceedings of the Wiswesser Line Notation meeting of the Army chemical information and data systems program, 6–7 October, 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. Edgewood Arsenal Special Publication EASP 400-8 (AD 665 397) Edgewood Arsenal, Maryland. 1968, 212 p.

172. Spialter, L. The atom connectivity matrix (ACM) and its characteristic polynomial (ACMCP): a new computer-oriented chemical nomenclature. J. Amer. Chem. Soc. 85(13):2012–2013. 1963.

173. Spialter, L. The atom connectivity matrix (ACM) and its characteristic polynomial (ACMCP). J. Chem. Doc. 4(4):261–269. 1964.

174. Spialter, L. The atom connectivity matrix characteristic polynomial (ACMCP) and its physico-geometric (topological) significance. J. Chem. Doc. 4(4):269–274. 1964.

175. Stobaugh, R. E., W. H. Powell, and R. J. Zalac. Systems for registering and naming polymers at Chemical Abstracts Service. (Papers presented

at 154th meeting, American Chemical Society, Division of Chemical Literature. Chicago, 1967.)

176. Stockton, F. G. Linearization and standardization of graphs; control of a factorial enumeration related to graph isomorphism. Shell Development Company. Tech. Progr. Rep. 2-68 (Project 34430: Techniques of Mathematical Analysis), Emeryville, California. 1968, 21 p.
177. Sussenguth, E. H. Use of tree structures for processing files. Association for Computing Machinery. Communications of the ACM. 6(5):272–279. 1963.
178. Sussenguth, E. H. A graph-theoretic algorithm for matching chemical structures. J. Chem. Doc. 5(1):36–43. 1965.
179. Sutherland, G. Dendral—a computer program for generating and filtering chemical structures. Stanford University. Artificial Intelligence Project. Memo 49, Stanford, California. 1967, 24 p.
180. Tate, F. A., and others. A mechanized registry of chemical compounds. (Paper presented at the 1965 Congress of the International Federation for Documentation. Washington, 1965.)
181. Tate, F. A. Handling chemical compounds in information systems. p. 285–309 In C. A. Cuadra, ed. American Documentation Institute Annual Review of Information Science and Technology. Vol 2. Interscience, New York. 1967.
182. Tate, F. A. Progress toward a computer-based chemical information system. Chem. Eng. News 45(4):78–90. 1967.
183. Tauber, S. J. Absolute configuration and chemical topology. U.S. Department of Commerce, National Bureau of Standards. J. Res. NBS, A. Physics and Chemistry. 67A(6):591–599. 1963.
184. Tauber, S. J. Digital handling of chemical structures and associated information. p. 206–216 In Association for Computer Machinery, Proceedings. ACM 20th National Conference, 1965.
185. Tauber, S. J., G. F. Fraction, and H. W. Hayward. Chemical structures as information representations, transformations, and calculations. p. 73–101 In B. F. Cheydleur, ed. Colloquium on Technical Preconditions for Retrieval Center Operations. Spartan Books, Washington. 1965.
*186. Tauber, S. J., and others. Algorithms for utilizing Hayward notations. U.S. Department of Commerce. National Bureau of Standards. Washington. [1965?].
187. Tauber, S. J., and others. Work toward chemical structure manipulation. U.S. Department of Commerce. National Bureau of Standards. Center for Computer Sciences and Technology. Report NBS 9007, [Washington]. 1965, 45 p.
188. Thompson, L. H., E. Hyde, and F. W. Matthews. Organic search and display using a connectivity matrix derived from Wiswesser notation. (Paper presented at 153rd National Meeting, American Chemical Society, Division of Chemical Literature. Miami Beach, 1967.)
189. Uchida, H., T. Kikuchi, and K. Hirayama. Mechanized retrieval system for organic compounds: an evaluation of the fragmentation code system. (Paper presented at the 33rd Conference of Federation International de Documentation and International Congress on Documentation, Tokyo, 1967.)

190. U.S. Army Chemical Information and Data System, status report. U.S. Army. Office of the Chief of Research and Development. Director of Army Technical Information. Report CIDS-1, 273 p. (AD 432 000) Army Research Office, Washington. 1964.

191. Uses of electronic computers in chemistry. (Summary of proceedings of Conference on Uses of Electronic Computers, Indiana University, Bloomington, Indiana, 1965.) National Academy of Sciences–National Research Council, Washington. 1967, 31 p.

192. Van Meter, C. T., S. D. Bedrosian, and D. Lefkovitz. Preliminary analysis of functional requirements for chemical information data systems. U.S. Army. Edgewood Arsenal. Chemical Research and Development Laboratories. Report CIDS-1, (AD 453 025) Edgewood Arsenal, Maryland. 1964, 65 p.

193. Van Meter, C. T., D. Lefkovitz, and S. D. Bedrosian. Comprehensive summary report on a proposed chemical information and data system. U.S. Army. Edgewood Arsenal. Chemical Research and Development Laboratories. Report CIDS-3, 2 Vols. (AD 479 679) Edgewood Arsenal, Maryland. 1965.

194. Van Meter, C. T., D. Lefkovitz, and S. D. Bedrosian. A study for a proposed chemical information and data system. U.S. Army. Edgewood Arsenal. Chemical Research and Development Laboratories. Report CIDS-2 (AD 477 110) Edgewood Arsenal, Maryland. 1965, 149 p.

*195. Van Meter, C. T. Chemical Abstracts systematic chemical nomenclature. University of Pennsylvania. Institute for Cooperative Research. Philadelphia. 1966, 17 p.

*196. Van Meter, C. T. Display of CIDS structures in standard form. University of Pennsylvania. Institute for Cooperative Research. Philadelphia. 1966, 21 p.

*197. Van Meter, C. T., D. Lefkovitz, and R. V. Powers. An experimental chemical information and data system. U.S. Army Edgewood Arsenal. Technical Support Directorate. Status Report CIDS-4 (AD 657 575) Edgewood Arsenal, Maryland. 1966, 223 p.

*198. Van Meter, C. T. Utilization of the Wiswesser notation in CIDS. p. 121–137 In J. P. Mitchell, ed. Proceedings of the Wiswesser Line Notation meeting of the Army chemical information and data systems program, 6–7 October, 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. Edgewood Arsenal Special Publication EASP 400-8, (AD 665 397) Edgewood Arsenal, Maryland. 1968, 212 p.

199. Vander Stouw, G. G., I. Naznitsky, and J. E. Rush. Procedures for converting systematic names of organic compounds into atom-bond connection tables. J. Chem. Doc. 7(3):165–169. 1967.

200. Vlannes, P. N., P. D. Olejar, and M. H. Weik. U.S. Army chemical information and data system (CIDS); some general aspects. (Paper presented at fifty-seventh annual meeting, American Institute of Chemical Engineers. Symposium on systems for estimation of physical properties. Boston, 1964.)

201. Vleduts, G. E. Concerning one system of classification and codification of organic reactions. Inform. Storage and Retrieval. 1(2-3):117–146. 1963.

*202. Vleduts, G. E., and G. R. Mischenko. Single algorithm for classification of reactions in organic chemistry. Inform. Sistemy. Moscow:88–106. 1964.

203. Weaver, G. A mathematical model for chemical cipher systems 1. University of Pennsylvania. Department of Linguistics. Analysis of Chemical Notations Project. Paper No. 2. 1968, 31 p.

204. Welch, J. T. Substructure searching of chemical compounds using Polish-type notation. J. Chem. Doc. 5(4):225–229. 1965.

*205. Whittingham, D. J., F. R. Wetsel, and H. L. Morgan. The computer-based subject index support system at Chemical Abstracts Service. J. Chem. Doc. 6(4):230–234. 1966.

206. Willard, J. R., and E. J. Malkiewich. A working system for retrieval of chemical structures, adaptable to pesticidal screening data. J. Chem. Doc. 4(4):211–217. 1964.

207. Wilson, W. J. An algorithmic approach to the problem of graph isomorphism determination. U.S. Army Missile Command. Redstone Scientific Information Center. Report RSIC-203, Redstone Arsenal, Alabama. 1964, 13 p.

208. Wilson, W. J., and J. B. Burger. A collection of algorithms for searching chemical compound structure analogs. U.S. Army Missile Command. Redstone Scientific Information Center. Report CIDS-3, (AD 460 819) Redstone Arsenal, Alabama. 1964, 21 p.

209. Wiswesser, W. J. A decoding study of the line-formula chemical notation, administered by the Reading Chemists' Club, 1961–1965. (Final report to the National Science Foundation.) (NSF-G15598) Reading, Pennsylvania, 1965.

210. Wiswesser, W. J. Computer-generated correlations among the organophosphorus pesticides. (Paper presented at 153rd National Meeting, American Chemical Society. Miami Beach, 1967.)

211. Wiswesser, W. J. The "Dot-plot" computer program. p. 103–116 In J. P. Mitchell, ed. Proceedings of the Wiswesser Line Notation meeting of the Army chemical information and data systems program, 6–7 October, 1967. U.S. Army Edgewood Arsenal. Technical Support Directorate. Edgewood Arsenal Special Publication EASP 400-8 (AD 665 397) Edgewood Arsenal, Maryland, 1968, 212 p.

*212. Wiswesser, W. J. Computer applications of the "WLN" (Wiswesser Line-Notation). U.S. Army. Plant Sciences Laboratories. Fort Detrick, Maryland. Technical Manuscript 490, Frederick, Maryland. 1969, 31 p.

213. Wurm, B. R. The discriminatory power of the biological terms of U.S. pharmaceutical patents for information-retrieval purposes. Proceedings of the 3rd Annual Meeting of the Committee for International Cooperation in Information Retrieval among Patent Offices (ICIREPAT), Vienna, September 1963, Reports No. 1 and 2, Spartan Books, Baltimore, Maryland. 1964, p. 277–305.

214. Zabriskie, K. H. Five-year plan for a computer-based chemical information system at the Chemical Abstracts Service. (Paper presented at the Chemists' Club Library Symposium on New Developments in Chemical Information Sources. New York, 1965.)

215. Zabriskie, K. H. Information retrieval operations at the Chemical Abstracts Service. p. 43–54 In B. F. Cheydleur, ed. Colloquium for Technical Preconditions for Retrieval Center Operations. Spartan Books, Washington. 1965

216. Ziegler, H. J. A new information system for organic reactions, J. Chem. Doc. 6(2):81–89. 1966.

INDEX

*