The Chemical Abstracts Service Chemical Registry System. VIII. Manual Registration

J. P. MOOSEMILLER,* A. W. RYAN, and R. E. STOBAUGH

Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210

Received February 4, 1980

The Chemical Abstracts Service (CAS) Chemical Registry System is a computer-based system that uniquely identifies chemical substances on the basis of structure and composition. For the small number of substances for which there is little or no structural description, special procedures have been developed. Analogous procedures are also necessary to process those substances that have completely described structural characteristics, but which lie beyond the bounds of the system. Examples of such substances are polypeptides, enzymes, "Colour Index" names where the dye structure has not been revealed, alloys, etc. The manual identification process and the combination of manual and machine processing necessary to record these substances in the CAS Chemical Registry System are described. Registration of compounds in the UVCB section of the TSCA Inventory is also discussed.

INTRODUCTION

The Chemical Abstracts Service (CAS) Chemical Registry System is a computer-based system that uniquely identifies chemical substances on the basis of their molecular structure. The descriptions of the system published to date[1-7] all have been aimed specifically at "machine registration" of chemical substances. This type of registration entails generating computer representations for chemical substances that can be described completely in structural terms such as atoms and bonds, and for certain specific types of substances that can be described only incompletely.[3] However, there are a small number of chemical substances for which there is little or no structural description available, or which cannot be adequately described in machine terms and thus exceed the limits of machine processing, but which are indexed in *Chemical Abstracts* or are processed by CAS for other organizations, and therefore must be recorded in the CAS Chemical Registry System. These substances must be identified and processed manually since their molecular structures cannot be represented by the unique, unambiguous connection tables in our current systems.

As of December 31, 1979, the CAS Manual File presently consists of 136 796 compounds. This represents about 3% of the total CAS Chemical Registry System data base which comprises 4 787 991 compounds. Statistics for the Manual File are categorized by chemical class in Table I.

This article presents the rationale for the necessity of manual registration. It describes in general terms the procedures used to verify and update manual registrations as well as some of the problem areas of the manual registration system. A historical account of CAS's efforts to decrease the number of chemical substances which must be manually registered is also presented.

BACKGROUND AND EARLY DEVELOPMENT

The initial computer-based system, known as Registry I, began operating late in 1964, when its major function was the support of substance registration for the published and computer-readable versions of *Chemical Biological Activities* (CBAC). It was recognized then that three categories of substances could not be machined registered: (1) substances for which complete structural representations could be determined, but which exceeded certain limitations of the computer operations; (2) substances for which only incomplete structural representations could be determined such that they were described insufficiently for input into the machine system; and (3) substances for which no structural descriptions could be determined, which had only nondescriptive names and which lacked the necessary information for machine registration.

Table I. Manual File Statistics by Category

Alloys	5,100
Enzymes	5,300
Elementary Particles	2,300
Colour Index Names	3,100
Name-only File	63,000
Carbon Molecular Formula File	31,500
Non-carbon Molecular Formula File	26,500

Comprising the first category of substances were typically polypeptides, polysaccharides, and polynucleotides, all of completely determined structure. Since the first version of the algorithm that generated the unique connection table was capable of handling substances containing no more than 150 nonhydrogen atoms, no structures exceeding this limit could be machine registered. For substances in the second category, the degree of incompleteness ranged from structures for which the location of only one substituent or one double bond was not specified to those for which only a molecular formula was known. Examples of the third category included naturally occurring substances for which no structural information or molecular formula was reported in the article being indexed.

The procedure for registration of these substances, whatever their category, involved some or all of the following steps: (a) preparation of conventional Registry Forms[2] during document analysis (a Registry Form was assigned to each specific substance identified during analysis and the name or names of that substance taken from the document, molecular formula if any, and whatever structural information was reported were recorded on the form); (b) manual comparison of the substance information on the Registry Forms with that contained in the master files of Registry Forms for manually registered substances ordered by name and by molecular formula; (c) manual assignment of a previously used CAS Registry Number if the substance had already been registered, or manual assignment of a new CAS Registry Number if the substance was new to the files; (d) entry via keyboard of the CAS Registry Number, trivial or trade name, CA Index Name, and molecular formula, if any, into appropriate computer-based index support systems.

0095-2338/80/1620-0083$01.00/0

The CAS Registry Numbers for these manually registered substances, including the check digits, were machine generated like those assigned to machine-registered substances. Initially, certain ranges of CAS Registry Numbers were set aside for manually registered substances, but this practice was later abandoned.

Substances Requiring Special Processing. The registration and indexing of natural products presented some unusual problems. It often happened that a structure and a name were proposed for a substance either when its isolation was first described in a publication or at a later date. The name usually was not systematic in the sense of describing the structure and generally was based on the source material. The question existed of how to handle registration of such substances. The structure could, of course, be machine-registered and a CA Index Name produced, all linked with the author-supplied name, but the structure might later be disproven and another one confirmed. In that case, the author-supplied name, which commonly would be used in the literature, would be associated with the manually assigned number, and a machine-assigned number(s) with the structural representation(s). At the time a structure was established for the substance, the manually assigned number was removed from use. All the names and references which were associated with the manually assigned number were transferred to the appropriate machine-assigned CAS Registry Number. After some time, however, this procedure became exceedingly cumbersome and difficult to administer and was discontinued in favor of the present techniques, which are described later.

Chemical substances for which only a name was available also presented handling problems, especially for identification, since there might be no way of knowing when two or more common names would actually refer to the same substance. As a rule, registration was performed only when the name was accompanied by sufficient characterizing information to assure identification.

Certain materials that had some substance-like aspects were intentionally not registered in the early Registry operations. Such materials were manufactured articles (e.g., gauze); plants and their components; animals and their components; a general class of serums, toxins, vaccines, and viruses; and generic derivatives (e.g., chloro derivatives of anthracene or glucose acetates). In the present registration process, as will be described later, this policy has been somewhat modified.

Registry II. When Registry II came into effect in 1968, major enhancements were made to the CAS Chemical Registry System to allow the machine registration of many compounds that were incompletely described. Some examples of these types of compounds were those containing substituent attachments at unknown positions, compounds where part of the structure was represented by only a molecular formula component, polymers which could be represented by a repeating unit, and compounds with unsaturation in an unknown position. Although these compounds can now be machine-registered, discussion of them is included below in the section describing the restrictions of the Registry III System because some rare variations of these types still exceed system limitations.

REGISTRY III AND ITS TOPOLOGICAL RESTRICTIONS

The present generation of the CAS Chemical Registry System, Registry III, came into existence in 1974. Registry III, like Registry II, represents modifications and enhancements to the algorithms of the CAS Chemical Registry System to allow machine registration of many more previously unregistrable compounds. Although the great majority of substances cited in the chemical literature have structures that

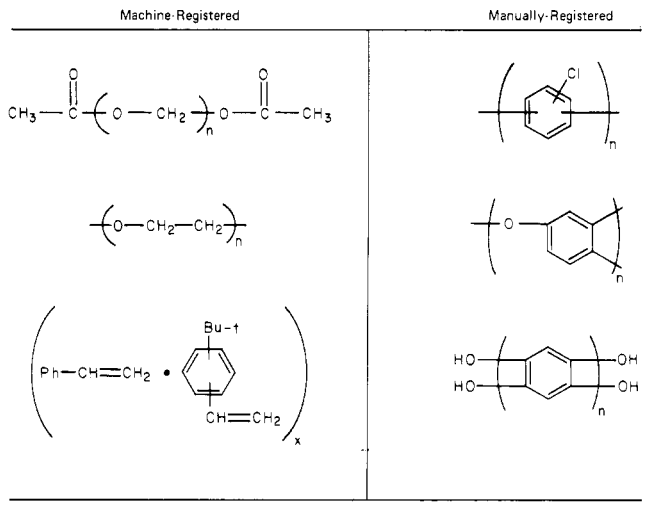

Figure 1. Examples of polymeric structures.

can be determined and registered by generating a unique computer-language description of their molecular structures via algorithms developed by CAS, there still exist certain unique chemical substances whose structures are not known, or which are of such size or indeterminate nature that a satisfactory machine representation cannot be made at this time. The types of substances that fall into this category, and consequently exceed the topological restrictions of Registry III, include the following.

Substances Whose Structures Are Unknown. Registry III identifies chemical substances on the basis of their molecular structure. Substances whose structural identities are not even partially known, but which are known to be specific chemical compounds (e.g., due to their isolation from natural sources and/or physical characterization), cannot be registered on a structural basis. This type of compound must be registered manually under Registry III. The basis for the registration or identification of the compound must then be its name as reported in the literature, whether of a systematic (containing structural information) or trivial nature. The term "trivial name" refers to a name which does not contain chemical structure information. Examples of this type of compound include natural products whose structures have not been elucidated or manufactured substances, known by their trade names, whose structures are not known or are concealed for commercial reasons. Substances of this type are registered manually according to the procedures outlined later in this article and are given machine-generated CAS Registry Numbers. They appear in the CAS Chemical Substance Index but have no Molecular Formula Index entry.

Substances Whose Structures Are Partially Known. Although many chemical substances for which CAS has identified partial structural information can be registered under Registry III, some cannot. For instance, substances that can be machine-registered include polymers which can be described in terms of one or more monomers and those which can be described in terms of a single structural repeating unit. A structural repeating unit is defined as a structural segment terminated at each end by an open bond, which is linearly repeated a large number of times. This is indicated by parentheses and a subscript "n" as shown in Figure 1. The open bonds may or may not be delimited by end groups. All these types of polymeric chemical substances, consisting of the following polymer classes—oligomers, homopolymers, telomers, copolymers, and simple structural repeating units—may be machine-registered.

Polymers with indefinite repeating units cannot be machine-registered because of limitations of the algorithm.

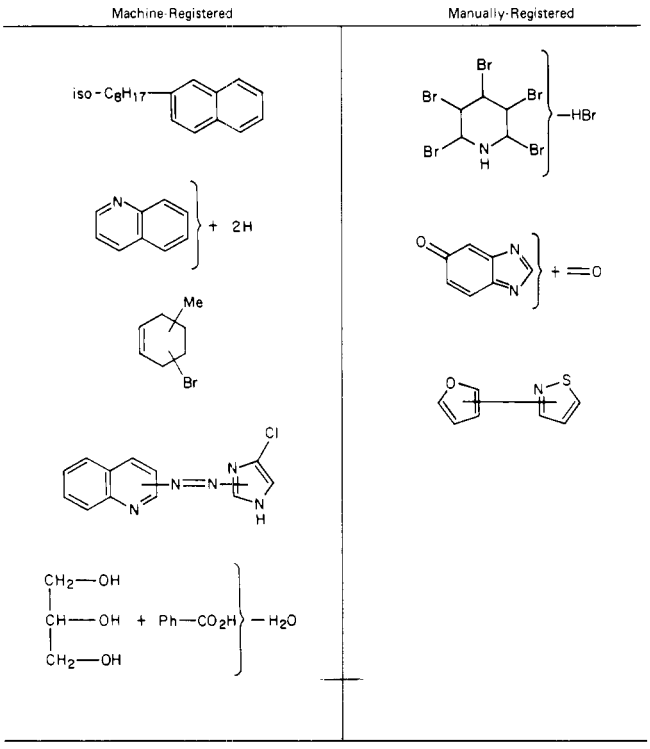

Figure 2. Incompletely defined compounds.

Likewise, repeating unit polymers which have no end groups but contain an odd number of connective bonds cannot be machine-registered. Polymeric substances described in terms of a repeating unit bounded by two end groups can be machine-registered, but those with repeating units of more than two end groups or an odd number of end groups cannot. Such substances are rarely described in the literature, but they do represent the limitations of Registry III algorithms. All of these examples are illustrated in Figure 1.

Another group of substances containing partially complete structures, which CAS refers to as incompletely described (ID) substances, falls into several classes, four of which can be machine-registered.[3] The first class encompasses compounds in which part of the structure can be expressed only as a molecular formula fragment. The second class comprises "hydro derivatives" in which saturation or unsaturation at an unknown position is indicated by the addition or subtraction of pairs of hydrogen atoms. Chemical substances having a substituent(s) attached at an unknown position constitute the third class of ID compounds. The fourth class of ID compounds are those which may be represented as having been formed by the reaction of two or more chemical substances with the loss of water (e.g., esters, amides, ethers, oximes, and hydrazones) where the multiplicity of possible reacting sites precludes an exact knowledge of the resulting chemical structure. Examples for these substance classes are shown in Figure 2.

In addition to these four classes, there are also certain types of ID substances that cannot be machine-registered. An ID compound represented by the addition or subtraction of atoms other than hydrogen or water cannot be machine-registered. Another limitation of Registry III prevents the machine-registration of a compound that contains two rings that are connected by indefinite attachments without any intervening atoms. However, polymers possessing incompletely described monomers can be machine-registered. Figure 2 illustrates some ID compounds, both machine- and manually-registrable types.

Certain types of indefinite ratio compounds cannot be machine-registered. For example, indefinite ratio esters of polybasic acids and/or polyalcohols must be manually registered, e.g., estradiol sulfate. Indefinite ratio esters of structural repeating units or of any ester of a homopolymer or copolymer must be registered manually, e.g., poly(ethylene glycol) phosphate and acrylic acid–butadiene polymer methyl ester. The same restriction also applies to indefinite ratio ethers of polyglycol structure repeating units and to indefinite ratio esters of mixed polyglycol copolymers.

Substances Whose Structures Exceed System Limitations. There are certain system limitations of Registry III with regard to the storage space allotted for the computer representation of an element of data. Most of these limits are so high as to be seldom or never encountered; e.g., the limits for the charge on an atom range from −127 to +126. There is also a limit of 64 abnormal valences and 64 rings for any machine-registrable fragment. Situations rarely arise where these limits are exceeded.

There are some storage limits of the current system, however, which make it impossible to machine-register certain types of chemical substances even though their structures are completely known. Registry III does not allow the machine input of components with greater than 253 nonhydrogen atoms. A class of compounds falling into this category is enzymes, some of whose structures are known but are too large for machine input. There are approximately 5300 manually registered compounds of this type in the CAS Chemical Registry System. Another limitation of Registry III that is occasionally reached is that the system allows a maximum of only 19 compound fragments for each single substance.

MANUAL FILE DESCRIPTION

The integrity of manually registered compounds is maintained through the use of hardcopy files which contain copies of Registry Forms.[2] These forms contain, whenever possible, CA Index names, synonyms, CAS Registry Numbers, molecular formulas, text descriptors, and structures for the manually registered compounds. The Registry Forms may also contain editorial notes providing additional information which might be needed to identify the compound. This information can consist of physical properties, method of preparation, or the general class to which the compound belongs (i.e., fluorescent dye). This information is useful when the same nonsystematic name is reported for several different chemical substances.

For ease of access, the Manual File has been divided into several smaller files. There are general files ordered by name and molecular formula, and specific files for alloys, enzymes, elementary particles, and dyes identified by "Colour Index" names. Statistics for these files, which are described below, are given in Table I.

Most manually registered compounds are found in the general file which has been subdivided into three additional files. The first is a name-only file for those substances that have no structural information and thus must be registered on the basis of a name reported in the literature. Trade names of commercially manufactured substances and trivial names of natural products are common examples. The second file is ordered on the basis of molecular formula for those compounds that contain carbon. Polymers are the most frequent kind of compounds found in this file. The third subfile of the general file is also ordered on molecular formula but consists of those compounds that do not contain carbon in their molecular formula. Line formulas of inorganic salts are most frequently found in this file. Separate files are maintained for manual registrations established specifically for federal regulatory agencies, as described later.

Manually registered compounds also have their corresponding machine representations. They have no structures

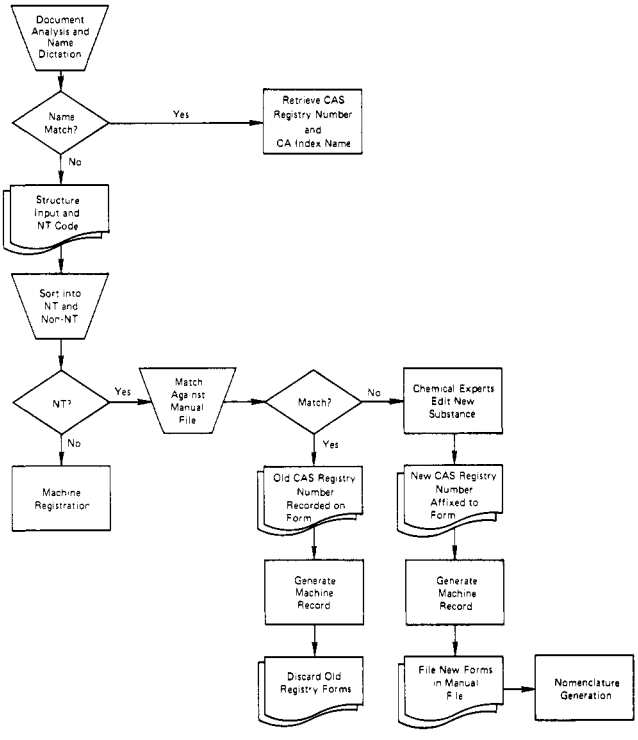

Figure 3. Manual registration process.

or connection tables, but they do have machine-generated CAS Registry Numbers, CA Index Names, and all other information that can be entered, including complete or partial molecular formulas when available. Special codes distinguish Manual File compounds from the rest of the substances in the CAS Registry Nomenclature File and the Registry Structure File.

THE MANUAL REGISTRATION PROCESS

The manual registration process explained below is outlined by the flowchart shown in Figure 3.

Document Analysis and Name Match. The document analyst decides which compounds in a particular journal article, report, or patent are of chemical interest. A temporary compound number is assigned to each chemical substance that the analyst chooses for indexing. At this point no distinction is made between machine-registrable compounds and those that will have to be manually registered.

The document analyst may initiate the identification of the substance by dictating a name for the candidate compound which will be used in a batch name-matching process. The name may be a trivial name used for the substance in the original document. A trade name for a commercially produced chemical substance, i.e., Dowfume W 85 for 1,2-dibromoethane, is an example of a trivial name. In the literature, natural products are often given trivial names which are derived from the source material, e.g., cinchonine.

The dictated name is subsequently keyboarded into the CAS computer-based index compilation system, and if it matches against a name in the Registry files, the CAS Registry Number is retrieved and associated with the compound number for that substance. Multiple name-match hits are available online for review by chemical experts. Batch name match is used for all chemical substances where applicable. It is especially important with regard to manual registry because many chemical substances that have been registered solely on the basis of a trivial name can be retrieved by the computer even though the compound was originally registered manually.

Structure Input. A chemical substance which does not name-match must go to structure input. At this point one of two procedures is followed: a chemist may draw a two-dimensional representation of the compound with appropriate stereochemical, coordination, or other information; or the appropriate CAS Registry Number for the compound may be added to the Registry Form, if that compound has been registered previously.

If the chemical substance falls into a class of compounds which must be registered manually, the chemist indicates this by coding NT (*n*o connection *t*able) on the Registry Form. The batch containing all Registry Forms for the document is then sent to Manual Registry where all forms with an NT code are removed. The rest are sent to machine registration. If the structure input chemist fails to recognize a candidate for Manual Registry, the structure will be rejected during the subsequent computer processing and be sent back to Manual Registry.

Matching against the Manual File. A Registry Form which has been coded NT is compared with similar compounds in the Manual File. If an unambiguous match can be made, the previously assigned CAS Registry Number is written on the Registry Form and then entered to the machine files as an index entry for the document in question. If an apparent match is of dubious nature, a chemical expert is consulted. Compounds which appear to be new to the Manual File are sent to chemical experts for review.

Editing by Chemical Experts. New candidates for the Manual File are examined for accuracy with respect to CAS chemical structure and naming conventions. The Manual File is again searched for possible matches. Reference books and additional internal files (e.g., the Natural Products File or the Drug Trade Name File) may be consulted for possible synonyms. If the Registry Form contains new information for a compound which is already in the Manual File, the file information is updated to reflect the change. A compound whose structure is elucidated in the course of several reports in the chemical literature (i.e, a natural product) may eventually be updated to a structural representation that can be machine-registered.

Machine Input. If the chemical substance appears to be new to the CAS Chemical Registry System, it is assigned a new machine-generated CAS Registry Number. With the exception of a structure connection table, the pertinent information is then entered into the machine record. As stated above, a special code is associated with the machine record to indicate that the compound has been registered manually.

Name Generation. Once a new compound has been successfully input into the system, it is still incomplete until a CA Index Name has been generated for it. This is done by a process in which compounds are separated by chemical classes and then named by nomenclature experts in the various chemical fields. There is no special provision for naming manually registered compounds.

Structure Update. As was mentioned earlier, the area of natural products often presents a problem in that a trivial name may be proposed for an isolable chemical substance whose definite structural entity has not yet been determined. However, the complete structural identity may be elucidated at a later date. Once a registration has been established for such a substance, CAS policy is to update the structural information associated with the corresponding CAS Registry Number rather than to create a new registration.

The process of structure updating occurs in much the same way as registering a new compound. A two-dimensional representation of the chemical substance is drawn on a Registry Form. However, instead of the information being associated with a new machine-generated CAS Registry Number, it is used to replace the incomplete or inaccurate data in the Registry Structure File for that previously established regis-

tration. In this way a natural product, or any chemical substance registered on the basis of a trivial name, can be updated to a more correct structural representation. If the updated structure can be represented by a computer-language connection table, it can be removed from the Manual File. Otherwise, the Manual File will be updated to reflect the new registration.

Occasionally, two or more registrations are established for the same chemical substance. This situation may arise from the appearance of the substance in the original literature in association with different trivial names and/or different levels of structural detail. In the natural product area the same compound may be isolated by various authors who propose dissimilar structures and assign different names to them. Similar situations may occur when commercial products or intermediates are initially identified only by their trade names. When the substances are subsequently shown to be synonymous, a cross-reference is established from the discontinued CAS Registry Number to the number which is retained. Cross-references of this type appear in the CAS Registry Handbook—Registry Number Update. Although the replaced CAS Registry Number is no longer valid, retrospective searches should take into account the fact that the number was valid in the past.

More rarely, a substance which has been registered solely on the basis of its trivial name is sometimes not found to be a unique chemical compound. In fact, the name may be generic and represent a broad class of compounds, or it may subsequently be indentified as a material which CAS does not classify as a chemical substance (e.g., rocks or mineral classes). The CAS Registry Numbers associated with such cases are declared invalid and appear in the Registry Number Update with the words "no longer in use".

MANUAL REGISTRATION FOR REGULATORY AGENCIES

Toxic Substances Control Act. In 1977 CAS was awarded a contract by the United States Environmental Protection Agency (EPA) for technical support of the Toxic Substances Control Act (TSCA). The purpose of TSCA was to produce an inventory of all chemical substances that are manufactured, imported, or processed for commercial purposes in the United States.[8] CAS agreed to process reports submitted by chemical manufacturers to EPA and to produce an inventory list of the chemicals for publication.

As part of this effort, CAS agreed to provide CAS Registry Numbers for all nonconfidential substances represented in the inventory. This included providing CAS Registry Numbers for reported substances that did not necessarily conform to the regular registration criteria for substances cited in the CA Chemical Substance Index.

Examples of substances that could not be represented by a definite chemical structure and which did not conform to regular CAS registration criteria were linseed oil, superphosphate, and blown castor oil. These types of compounds corresponded to those chemical substances which must be manually registered in the CAS Chemical Registry System. Within the context of the EPA TSCA Inventory they appeared in a special section—the Chemical Substances of *U*nknown or *V*ariable Composition, *C*omplex Reaction Products, and *B*iological Materials (UVCB) Section. The nature of the UVCB substances caused them to fall between the specific chemical compounds appearing in the CAS Chemical Substance Index and the general entries appearing in the CAS General Subject Index.

The UVCB substances that were more specific often corresponded to chemically undefined or indefinite derivatives of specific chemical substances as registered by CAS. For the purposes of producing *Chemical Abstracts*, such compounds were indexed and registered at the specific compound entry, and information was added to the index which described the indefinite derivative. As a typical example, chlorinated *o*-phenylphenol would be registered as *o*-phenylphenol with the term "chlorinated" added in the index. For the production of the EPA TSCA Inventory, this additional information was included as part of the data corresponding to the CA Index Name. These compounds were then manually registered and were referred to as "generic registrations".

The UVCB substances which more closely resembled the general subject headings published in indexes to *Chemical Abstracts* were also input to the substance-handling system and named like the general subject headings. A typical example, "castor oil, hydrogenated", is a modified CA general subject heading and was therefore treated as such. These registrations also included certain biological materials (e.g., *Saccharomyces cerevisiae*) which were commercially produced as defined by the EPA. It was also necessary for these registrations to be manual since there was no corresponding specific structure.

UVCB Machine Records. The new CAS Registry Numbers resulting from work done under the TSCA contract were consistent with and contiguous to those generated for chemical substances registered for citation in CA publications. However, the UVCB manual registrations were given separate codes to identify them and to keep them separate from the registrations generated by CAS as part of producing its routine publications.

As of December 31, 1979, there were a total of 3631 generic registrations and 9060 general subject-like registrations added to the CAS Chemical Registry System. These figures were not included in the Manual File total given earlier, but are included in the total CAS Chemical Registry System registration figure which is 4 787 991.

Publication of UVCB Registrations. Whenever the CAS Registry Numbers for UVCB chemical substances have appeared in TSCA publications, asterisks were appended to them. This is to indicate that these registrations were specially created, and that their CAS Registry Numbers will not be found in routine CAS abstract and index services. The only CA publication where these registrations appear is the Registry Handbook, where these CAS Registry Numbers also have an asterisk appended to them.

SUMMARY

The CAS Chemical Registry System uniquely identifies chemical substances on a structural level by deriving computer representations of the molecules' atoms and bonds. Chemical substances whose structures are not known, or which cannot be adequately described in machine terms, must be registered manually. In order to recognize duplicates and maintain the integrity of manually registered compounds, a hardcopy Manual File must be routinely referenced. This Manual File consists of Registry Forms containing the most complete description of the compound available. When the analysis of a document results in the indexing of a chemical substance that cannot be given a machine representation, the Manual File is consulted to determine if the compound is old or new to the CAS Chemical Registry System. When the CAS Registry Number and CA Index Name are determined for such a chemical substance, they are then input into the computer records that are used to produce CAS publications and services.

ACKNOWLEDGMENT

The CAS Chemical Registry System was developed with the substantial support of the National Science Foundation. Chemical Abstracts Service, a division of the American

Chemical Society, gratefully acknowledges this support.

REFERENCES AND NOTES

(1) Morgan, H. L. "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service", *J. Chem. Doc.* **1965**, *5*, 107.
(2) Leiter, Jr., D. P.; Morgan, H. L.; Stobaugh, R. E. "Installation and Operation of a Registry for Chemical Compounds", *J. Chem. Doc.* **1965**, *5*, 238-242.
(3) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Chemical Registry System. I. General Design", *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111-121.
(4) Freeland, R. G.; Funk, S. A.; O'Korn, L. J.; Wilson, G. A. "The Chemical Abstracts Service Chemical Registry System. II. Augmented Connectivity Molecular Formula", *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 94-97.
(5) Blackwood, J. E.; Elliott, P. S.; Stobaugh, R. E.; Watson C. E. "The Chemical Abstracts Service Chemical Registry System. III. Stereochemistry", *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 3-8.
(6) Vander Stouw, G. G.; Gustafson, C.; Rule, J. D.; Watson, C. E. "The Chemical Abstracts Service Chemical Registry System. IV. Use of the Registry System to Support the Preparation of Index Nomenclature", *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 213-218.
(7) Zamora, A.; Dayton, D. L. "The Chemical Abstracts Service Chemical Registry System. V. Structure Input and Editing", *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 219-222.
(8) "Inventory Reporting Regulations (40 CFR 710)", *Fed. Regist.* **1977** (Dec 23), *42* (No. 247).

Computer-Assisted Synthetic Analysis at Merck†

PETER GUND, EDWARD J. J. GRABOWSKI, DALE R. HOFF, and GRAHAM M. SMITH*

Merck Sharp & Dohme Research Laboratories, Rahway, New Jersey 07065

JOSEPH D. ANDOSE* and JOSEPH B. RHODES

Management Information Systems, Merck & Co., Inc., Rahway, New Jersey 07065

W. TODD WIPKE

Board of Studies in Chemistry, University of California, Santa Cruz, California 95060

Received January 24, 1980

The Simulation and Evaluation of Chemical Synthesis (SECS) program has been implemented at Merck and has been evaluated by approximately 50 synthetic chemists. The results of this evaluation are summarized, highlighted by several examples of SECS analyses, and future plans for the program at Merck are discussed. The most critical problem which must be solved is that of developing a practical data base of synthetic reactions.

In the decade since the first published description of a computer program capable of deriving synthetic routes to complex molecules,[1] the field has grown into a lively discipline.[2-4] Nevertheless, it cannot yet be said that computer-assisted synthesis is an accepted research tool for many practicing synthetic chemists.

It is easy to see the appeal of computer aids to synthesis. Organic synthesis is complex; in principle, several thousand reactions could be applied in several steps to millions of potential starting materials in order to prepare a desired product.[2,3] The computer has the capability to extend the chemists's perception and memory, and to systematize and organize his synthetic knowledge.

While Merck has for a long time recognized the computer's potential in this area,[5] the decade was half over before the field was sufficiently advanced for an in-house effort to be considered. When a computer-assisted synthesis project was begun, a number of major design decisions were quickly made. To assure data security and to encourage optimal use, we wanted to run the program in-house. We wished to involve the synthetic chemists directly in the analyses, which required interactive time-shared program operation on graphics terminals remote from the computer. This also meant that synthetic analyses would not be run exclusively as a scientific information service. Finally, by using the same graphics terminals for computer-assisted synthesis and for the Merck Molecular Modeling Project,[7] we could make more efficient use of our resources. A brief survey of extant programs indicated that the Simulation and Evaluation of Chemical Synthesis (SECS) program[6,8,9] was most suited to Merck's needs.

Today we are running SECS on our corporate IBM computer, with the program accessed by four graphics terminals in three research laboratories (Rahway, New Jersey; West Point, Pennsylvania; and Montreal, Canada). About 60 chemists have been checked out on running SECS analyses, and the program is used almost daily. About 20 volunteer chemists have contributed chemistry for the program, and two chemists are adding chemistry to the file.

At this stage in the program's development, it is appropriate to report our experiences in implementing, evaluating, and enhancing the SECS program at Merck, and to assess progress toward the ultimate goal of providing a routinely useful aid for the practicing synthetic chemist.

IMPLEMENTATION OF SECS AT MERCK

This section gives a brief discussion of how SECS was converted to run in the Merck environment.[10]

In October of 1974, one of us (W.T.W.) made available version 1.0[11] of the program consisting of approximately 25000 lines of code written for a Digital Equipment Corp. (DEC) PDP-10 computer (ca. 20000 lines of FORTRAN written for

†Presented in part at the Symposium on Computer Assisted Drug Design at the 177th National Meeting of the American Chemical Society, Honolulu, Hawaii, April 1979. A more detailed version of this paper will be published in the Proceedings of this meeting.

0095-2338/80/1620-0088$01.00/0