answer which appears to provide the best balance between reduction of noise and loss of significant information.

The use of a save tape is illustrated in Figures 8 and 9. The patents on polymers made by Ziegler catalysis amounted to 306 and were stored on tape (Figure 8). Subsequently a question was asked to determine in how many of these transition-metal oxides figured in the catalysis. Of the 27 retrieved, a review showed that 9 were of interest.

When, as in this instance, related questions are saved on the same tape, it is advisable to repeat the listing of the original descriptors in searching the save tape. When other questions are not related or when only one question is saved, such relisting is unnecessary.

We have been curious to compare the performance of the computer-searched IFI Index with that of other indexes we have been using and continue to use. In one instance, patent searchers in Washington found, by using the U. S. Patent Office classification, four patents, three of which issued since 1950 having also been retrieved by use of the IFI Index. By an IFI search 105 patents were retrieved. From this number 18 of possible interest were selected by checking principal claims, which serve as abstracts. These 18 patents included the three post-1950 patents retrieved by search in the U. S. Patent Office. A search in other indexes available at Gulf Research & Development Company's Research Center had furnished a list of patents whose 1950+ portion had all been retrieved by the IFI search.

We have mentioned briefly above supplementing computer searching of the IFI Index with hand searching of other indexes. For example, in an exhaustive search, we list the Patent Office classes to which the patents retrieved in the IFI search have been assigned. This list may then serve as a guide to the searcher using the collection of classified patents in Washington. The inverse process, recalling the IFI descriptor list of a known pertinent patent, has already been mentioned.

A problem still to be solved is how best to reduce the number of patents printed out by the computer to the text which is to be read by the person or persons requesting the information. We assume here that the list has been pared to a reasonably low number by reframing questions. Some people who request searches are willing to work with the raw computer output themselves. In other instances, a member of the Patent Information staff does this task. In general, a person can screen 30–60 patents per hour; thus, this remains the expensive phase of searching.

Whoever screens the output may advantageously begin by referring to the IFI books of representative claims, mentioned above and illustrated in Figure 4. We keep these books in an air-conditioned room with convenient work space. We feel that, except in state-of-the-art searches, it is reasonable to limit the number of patents to be screened to 100. When the output exceeds that number, the question should be rephrased to obtain a less noisy output.

Copies of many of those patents whose full text the searcher wishes to read are on file in the same room. *All* patents from 1959 on in the IFI Index are available on microfilm, and we have a reader-printer in the search room so that these microfilmed copies can be conveniently read and, if desired, printed. Thus, we have tried to facilitate the use of the computer-search output by phrasing questions to obtain a reasonable signal-to-noise ratio and by putting all the equipment for screening the output into one place.

Installation and Operation of a Registry for Chemical Compounds*

DON P. LEITER, Jr., HARRY L. MORGAN, and ROBERT E. STOBAUGH
Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received July 7, 1965

Since 1958 the Chemical Abstracts Service has been working toward establishing a computer-based system for handling chemical information.** Briefly, the concept of the CAS system consists of sets of special subject files in the following categories: (1) physical properties, (2) chemical reactivities, (3) biochemical activities, and (4) applications. With the importance of compounds in correlation studies, and the need to interrelate compounds and the huge collections of chemical and other data, a highly developed subsystem, called the Registry System, for handling compounds must be the first step in the actual operation of an over-all computer-based service. The Registry System will include files of compounds interconnected with files of associated data that permit identifying the compounds and retrieving them from the files.

The process of registration or entry into the system includes the assignment of a unique machine address, called a Registry Number, to each compound which is new to the system or the retrieval of the previously assigned Registry Number from the files. It is intended that ultimately the system will include a record of every chemical substance reported in the literature and iden-

* Presented before the Division of Chemical Literature, Symposium on Current Applications of Mechanized Procedures for Handling Chemical Information, 149th National Meeting of the American Chemical Society, Detroit, Mich., April 1965.

** CAS is pleased to acknowledge the support of the National Science Foundation during all of this work and of the National Institutes of Health during part of the work.

tification of all useful published literature bearing on each substance.

A mechanized chemical compound Registry System is needed for a number of reasons, important among which are:

1. The traditional means of providing access to files of chemical compounds is through chemical nomenclature. However, nomenclature is not an adequate means of identifying and arranging compounds. The rules are too complex, too numerous, too inconsistent, and too incomplete for the average chemist to generate the systematic name. Thus the process of naming compounds must be carried on by specialists; this process is slow and expensive. Current-day pressures require a more rapid system.
2. Chemical compounds and materials occupy a substantial portion of the chemical literature; they relate, for example, to an estimated 85% of the subject entries in current *Chemical Abstracts* volume subject indexes.
3. Chemical science and technology needs a high-speed, automated means for ascertaining the location in the literature of information about compounds and for determining whether a compound is new or has been produced or identified previously. Such information will simplify the planning of research and development projects.
4. Efficient development of a national chemical information network depends on the early establishment of the Registry System so that interested organizations can pursue systems development in a coordinated manner assured by the availability of machine processable records of a body of important chemical materials and their literature references.

The CAS Registry System depends upon the identification of the structural diagram. The system is computer-based. In planning the operation, careful consideration has been given to the balance between the human and computer assignments so as to achieve speed as well as economical operation. The system has been designed to permit file buildup and useful operations *now* while providing for facile integration of improvements in elements which are still in various stages of research both at CAS and at other research centers. As improvements come along they can be incorporated into the system and utilized without re-analyzing the data that have already been filed. As experience is gained in operating the system, as research continues, and as improvements in equipment are made, the balance between human and computer assignments will shift toward increased computer involvement.

The versatility of the computer in organizing and reorganizing the files makes possible increased flexibility in use of the files and thus ensures a long useful life for the stored information. These factors in turn justify substantially more analytical effort than in preparing traditional printed information tools. The CAS Registry System has been designed to take full advantage of the computer's versatility and power. Staff requirements have been carefully considered in developing the system design.

Initially, the Registry will be confined to individual organic compounds for which a full, conventional, two-dimensional structural diagram can be drawn. Nonorganic compounds, polymers, and compounds with partially known structures initially will be omitted, although as research and development solve the problems of handling these substances, they will be included.

The following factors were involved in establishing the Registry System:

1. The CAS Input Bank is estimated to include at the present time 2.5 to 3 million chemical compounds. Most of these are organic compounds.
2. For the 1965 issues (Volumes 62, 63) of *Chemical Abstracts* (200,000 abstracts estimated), there will be about 1.3 million subject index entries, some 85% of which will deal with chemical compounds and materials. Thus, if all of the data related to a given compound are to be associated with that compound, and if it is assumed that the information should be entered into the computer files as quickly as the information is available from the indexer (analyst), then about 1.1 million items of compound- and material-oriented information would be processed through the Registry System this year.
3. This year about 75,000 compounds will be reported in the published literature for the first time.
4. Approximately 400,000 two-dimensional structural diagrams will be prepared this year in developing the CA Subject and Formula Indexes. The 400,000 structures do not correspond to 400,000 different compounds; rather the structures correspond to 400,000 occurrences of the compounds in the literature for which two-dimensional diagrams must be prepared as part of the identification and naming process for the CA Subject Index.

Since it is intended that the computer file will eventually include all compounds, with approximately 3 million compounds already known, we must be as specific as possible. For this reason, we assign a registry number to each compound. For example, (1) each salt has a different number; (2) a racemic mixture receives a number; (3) an ion receives a number when only the ion is identified; (4) a compound with unspecified stereochemistry is given a registry number different from the one in which the stereochemistry is specified.

The machine technique has not yet been extended to all types of structural detail, but techniques and computer programs are complete for generating the unique description for the two-dimensional projection of fully known nonpolymeric chemical structures. The third dimension is presently handled by the addition of conventional stereochemical descriptors which are supplied from the original publication by the chemist who prepares the structural diagram for input to the system.

Chemical-Biological Activities (CBAC) is the first special subject service at CAS based on the computer. The Registry System is an integral part of the CBAC publication process. The CBAC operational flow, illustrated in Figure 1, proceeds as follows.

Journal Acquisition. The journals acquired by the CAS library are used for CBAC, CT, and CA, as appropriate. A paper selected for CA may also be marked for inclusion in CT and CBAC. CT and CBAC provide special coverage within the full range of CA coverage. Thus, the contents of CT are limited by the selection of journals, and CBAC coverage is limited by journal and by subject definition. CA includes everything that is handled by CT and CBAC.

Digest and Structure Preparation. The analyst carefully reviews the original paper and then prepares a digest and an abstract. Thus, a single processing provides both

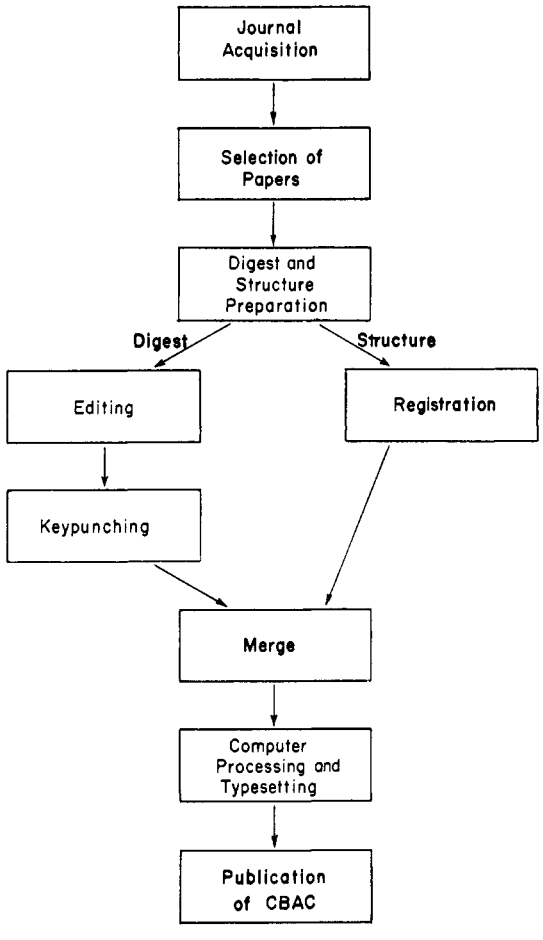

Figure 1. Operation flow for *Chemical-Biological Activities* (CBAC).

the CBAC digest and the CA abstract. While the digest and the abstract cover generally the same subject matter, the digest and the abstract differ in that the language used in CBAC is much more restricted than that used in CA.

Compounds that cannot be manually associated with a Registry Number are then structured.

After the digests and structures are prepared, the structures proceed through the registration process and the digest is edited, keypunched, and entered into the computer.

Following the registration and editing steps, the CBAC information is merged in the computer, processed, and composed for printing. Appropriate structural formulas are added to the copy produced by the computer, and the pages are sent to the printer for printing and distribution.

The Registry System is outlined in Figure 2.

At present, we are clerically preparing connection tables, and these are keypunched and fed into the computer for processing. When a tape-generating typewriter, such as the Army Chemical Typewriter developed at Walter Reed Army Institute of Research, becomes available, we plan to use such a device for input also. For typewriter input, the computer would be programmed to generate the connection table automatically from the typed record. We believe, however, that there will always be some structures that will be processed by manually generated

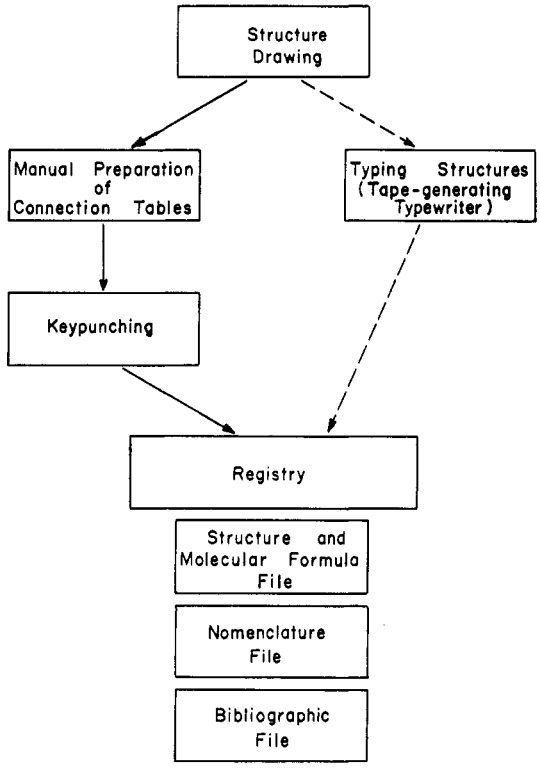

Figure 2. Operational flow for the Registry System.

connection tables. Registry is made up of three files: (1) structure and molform, (2) nomenclature, and (3) bibliographic.

CAS structure-drawing operations have been modified so that the registry form now includes all of the data necessary for the Registry System, as well as the information required for CA indexing operations. This form is shown on Figure 3.

A separate form is used for each structure. This form includes the Temporary Identification (TID) Number in the upper left-hand corner. This number connects the compound to the source document. A TID number identifies each compound; the number is unambiguous but it is not unique.

The names which appear in the original paper are recorded on the form by the analyst. The Registry Number is either provided as a consequence of the registration process, or it is supplied directly from a list of assigned numbers by the document analyst. (The method for generating this list is described later.) When the Registry Number cannot be quickly determined by the analyst, a structural diagram and molecular formula for the compound are placed on the form. A connection table is then prepared clerically. The information on the form is then keypunched and entered into the system, where registration occurs and the files are updated.

The machine record of each compound consists of a tabular record, called a Connection Table, which identifies each nonhydrogen atom and its interconnections within the molecule.

In manually preparing the connection table, the clerk randomly numbers each atom in the structure and proceeds to prepare the table. The numbering is not important so long as each nonhydrogen atom receives a different

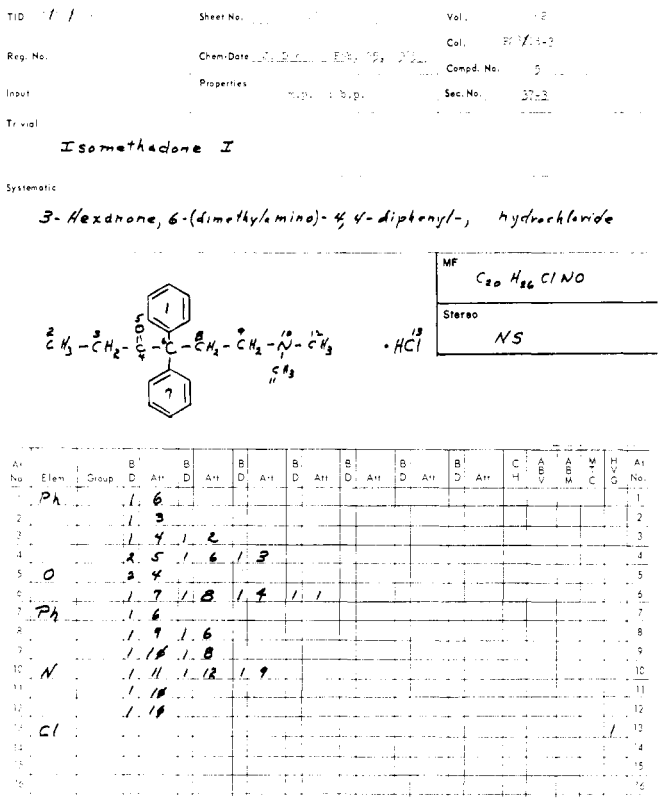

Figure 3. Example of Registry Form.

number. The table (Figure 3) shows the bonding; for example, No. 4 has a double bond to 5, a single bond to 6, and a single bond to 3. Of the nonhydrogen only the noncarbon elements are individually identified in the table. This example shows a Ph for phenyl group in positions 1 and 7; the computer expands this to the six corresponding ranks of the connection table. A carbon atom is identified in the element column only when it is the final atom in the table. The number of hydrogens must be indicated for any noncarbon element to which hydrogens are attached, in the final column of the table.

As entered into the computer, the description of the compound is unambiguous, but non-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 lexicographically in the file. Thus, the actual registration amounts to the merging of a list of additions with the registry file. The means of deriving the unique form of the connection table for registration is described in a recent paper (H. L. Morgan, *J. Chem. Doc.*, **5**, 107 (1965). This is an important aspect of the system, for the file will include in time over 3,000,000 compounds. Such a file must not include unrecognized synonymous entries; each compound must have a single representation in the file. Without this characteristic, the file could balloon to unmanageable proportions. Moreover, this characteristic—one entry and only one entry per compound—is necessary to assure the meaning of either a productive or a negative file search.

One means of simplifying the registration operation actually derives from the operation itself. Those compounds which are frequently reported in the literature are arranged by the computer into an index of names, which shows the corresponding registry numbers. Instead of drawing the structure for such compounds, the analyst simply looks up the number, and the bibliographic and nomenclature files are updated on this basis. Another similar list would index the large, complex structures that are expensive to prepare. These indexes we refer to as Desktop Analysis Tools. Associated with the Analysis Tools is what we call the Registry Handbook. The Handbook will include the structural diagram, the molecular formula, and the Registry Number for each compound in the file. The Handbook will be arranged in number order. Though the index for the full Handbook will include all compound names, it is not intended initially that the text of the Handbook will include nomenclature data because of the size of the updating problem. Another means of simplifying the registration will be the ability to register solely on the basis of well-known trivial names.

The Registry System must include a vigorous error avoidance effort if it is to continue to be useful. Some of the aspects of the purification routines are worthy of mention here. Such routines must handle several types of potential errors. Thus errors in the system could arise: (1) if the structural diagram as prepared by the chemist were in full accord with the rules of structure drawing, but the diagram did not correspond to the reported name (examples are shown in Figure 4); (2) if the structural diagram did not correspond to the connection table recorded in the computer because of an error in clerical transcription, but the editing routines did not find the error; (3) if there were a flaw in the algorithm for the generation of the unique table; (4) if an error existed in the programs that generate the machine representation and the unique table, or in the programs that assign the registry number and maintain the structure file.

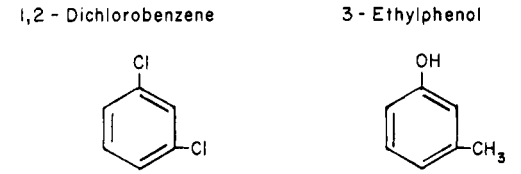

Figure 4. Structures are in full accord with the rules of structuring but do not correspond to the compounds named.

In order to constantly check for errors and to monitor the entire registry operation for "problem" areas we have built into the system permanent methods for detection and control of errors.

A. The first method by which errors are detected and actually barred from the system is by rigorous machine checking of the input structures. These checks of the molecular formula include valence and syntax.

B. The molecular formula is then computed from the connection table and checked against the formula computed by the chemist. When the values do not agree, the entry is rejected.

C. The second permanent method for detecting errors is performed as part of the maintenance of the "All Name File" and as such is a by-product of work necessary to the total information system. The All Name File is a computer store of every name in the system. Associated with each name is the registry number assigned to the corresponding structure. Each time this file is updated, the computer program will list as potential errors those

names which have been associated with two different registry numbers. These are, however, only potential errors because of the ambiguity of some chemical names.

D. A special index of molecular formulas and associated registry numbers will be produced by the computer at set intervals. A random sampling will be made of those sets of compounds which have the same molecular formulas. A chemist will compare the structures and the nomenclature to identify any structures which represent the same compound but which may have been assigned a different registry number.

Temporary Methods for Error Detection. In order to ensure adequate error detection and control during the early stages of production, temporary checking procedures have been instituted. Once the system has functioned smoothly in full operation, these procedures will be phased out, though each time the system is modified some variant of these checks will be reinstituted.

A. A method which has been used to test both the algorithm and the computer programs is the reprocessing by the computer of structures already registered. The processing is accomplished by a computer program which converts each connection table in the master file back to an expanded form of the table of the type which is input to the computer. (Note that in the resulting expanded table the numbering of the atoms is generally different from the numbering of the atoms in the unique form of the connection table and is also generally different from the numbering assigned at the time the compound was initially input to the system.) The computer registration process is then rerun and the results are machine checked to be certain that the resulting master file is identical with the one before the regeneration. If the two files are not identical, an error has been detected in the programs and/or the algorithm. This check device is no longer useful on the present system, for essentially all of the errors which can be uncovered by this technique have already been corrected. However, later alterations or improvements in the algorithm and programs will be checked in the same manner.

B. A major problem during the early stages of production is the training of the clerks who prepare the connection tables and the chemists who draw the structural diagrams. In order to reduce the potential error caused by incomplete training, a rigorous editing of the output of both the chemists and clerks has been instituted. In the case of the chemist, this takes the form of a review by a senior chemist to ensure correspondence between the name and the structure. In the case of the clerks, a second clerk edits the completed connection tables for errors prior to keypunching. A random sampling and checking operation is a standard part of the structure-drawing operation.

In addition to supporting the generation of CBAC, the Registry System will also routinely register each of the new entries for the Index of Ring Systems which is part of the CA Subject Index.

A Computer-Based Source Inventory of *Chemical Abstracts**

PETER M. BERNAYS, KENNETH L. COE, and JAMES L. WOOD
Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received July 21, 1965

One of the control measures recently instituted by Chemical Abstracts Service was the development of a computer-based inventory of abstracts. This step was taken in order to gain better insight into and control of production operations. The benefits derived from the inventory, and the uses to which it can be put, are discussed in this paper.

In the publication of *Chemical Titles*, CT (over 80,000 titles in 1964), the bibliographic data for each of the cited papers is stored in computer form. These data include the title of the paper in English (translated wherever necessary), the authors, the title of the journal stored as a four-letter code with the volume number, page numbers, and, where necessary, the issue number corresponding to the paper. The four-letter journal codes, called Coden, are the same as those adopted by the American Society for Testing and Materials.

When the publication of CT started in 1961, the median prepublication period (the period between the issuance of a primary journal and the appearance of a corresponding abstract in CA) was well over 7 months. By the end of 1964 this time period had been reduced to 3.8 months. These figures show why CT was originally set up in parallel with the publication of *Chemical Abstracts* and why these operations were integrated toward the end of 1964. During the three intervening years, processes were overhauled

* Presented before the Division of Chemical Literature, 149th National Meeting of the American Chemical Society, Detroit, Mich., April 1965.