Quality Control and Auditing Procedures in The Chemical Abstracts Service Compound Registry*

D. P. LEITER, Jr. and H. L. MORGAN**
Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received October 3, 1966

This paper describes quality control and auditing procedures which have been, or are about to be, instituted in the CAS Compound Registry System; the Registry System is briefly described.

Since early 1965, CAS has been developing an experimental Chemical Compound Registry System through a contract with the National Science Foundation acting on behalf of itself, the Department of Defense, the National Institutes of Health, the Food and Drug Administration, and the National Library of Medicine (*1*). The major purpose of the registration process is to determine whether a particular chemical structure is new to the System or has already been processed, and on the basis of this determination to assign a unique Registry Number to every chemical structure. The Registry Number is the thread that ties together all information associated with a particular compound throughout the developing CAS computer-based chemical information systems. Thus, through the registration process, CAS will be able to provide correlative searches of many files with assurance that all information on file for a particular compound has been located.

Presently the Registry System is building up three principal computer files—namely, (1) the structure file, containing the Registry Number and the computer representation of the structural formula for each registered compound (2) the Bibliography File, containing the Registry Number, the *CA* index nomenclature, the molecular formula, and the bibliographic citations for the compounds; and (3) the Nomenclature File, containing the Registry Number, the molecular formula, and all chemical nomenclature thus far input to the system.

These Registry Files are interlinked by the Registry Number and in addition are linked through the bibliographic references or through the Registry Number to other computer and manual information files at CAS (Figure 1).

For example, digests for *Chemical-Biological Activities (CBAC)* are stored in computer form for searching and are linked to the Registry files by the Registry Number of the compounds appearing in the digests. Also the *CA* issue Subject (Keyword) Index, Author Index, Patent Index, and the volume Subject and Formula Indexes, which are all in various stages of mechanization, are linked to the Registry files via a combination of *CA* references and/or Registry Numbers. Thus, searches carried out on one set of files or indexes can result in subsequent correlative searches and/or retrieval based on another set of files or indexes.

In order to maintain the requisite high degree of accuracy in the Registry Files, and at the same time reduce error-prone human operations to a minimum, CAS has developed and is continuing to develop computer-based error control and auditing procedures for all data input to the system. The discussion of these procedures here falls into two parts: (1) the procedures for structural data; and (2) those for references and nomenclature. Since the basis of chemical identity in the Registry System is the structure of a compound, our computer-based error-detection procedures have so far concentrated on structures. However, we are now developing similar computer-based procedures for nonstructural data.

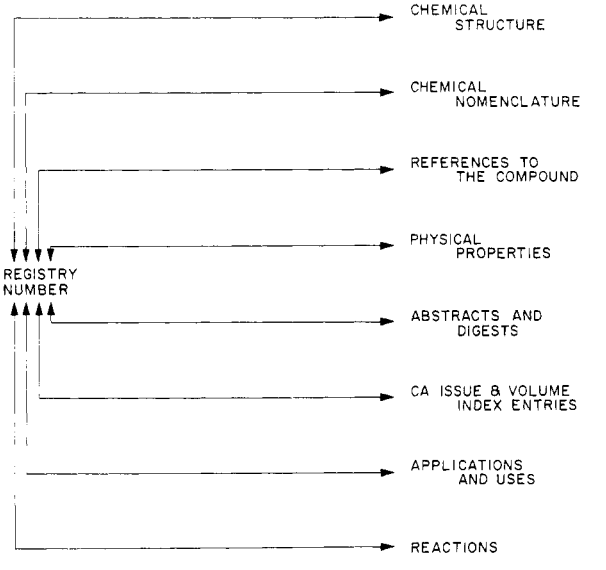

Figure 1. The Registry Number of a compound ties together information on that compound in many computer and/or manual files at CAS.

* Presented before the Division of Chemical Literature, 152nd National Meeting of the American Chemical Society, New York, N. Y., Sept. 15, 1966.

** Present address: IBM Corp., 1000 Westchester Ave., Harrison, N. Y.

COMPUTER CHECKS OF STRUCTURES

Inasmuch as the basis of registration is the computer representation of a compound's structure, it is extremely important that errors introduced by the chemist in drawing the structure or by the clerk in keyboarding the data for input be detected and rigorously barred from the files. The use of the computer for the detection of errors in chemical structures is not so difficult as one might at first imagine, for two important tools are available for the program to use. The first is the high degree of redundancy CAS has imposed on the input structural representation; the second are the well-established rules of chemical valence. Using primarily these two tools, the computer program is able to find most of the simple errors which are, of course, the most common in large-volume operations.

The processing of a structure for registration proceeds as follows. Using a standard, preprinted worksheet called a Registry Form, a chemist draws the conventional structural diagram for the compound and records the corresponding molecular formula. He also records, by means of conventional descriptors, any stereochemical detail for the compound. The Registry Form is then turned over to a clerk-typist, who numbers each nonhydrogen atom in the structure, and then keyboards a "connection table" for input to the computer. This table, which is completely equivalent in information content to the hand-drawn two-dimensional structure, lists each nonhydrogen atom by number and by elemental symbol and shows how each nonhydrogen atom is connected to other nonhydrogen atoms in the molecule. Such a connection table is illustrated in Figure 2, where the first rank can be interpreted to read "atom number 1, a nitrogen atom, is bonded with a double bond to atom number 8 and with a single bond to atom number 2." (Note that for actual registration, the connection table in Figure 2 would not be written out, but would be keyboarded directly for input to the computer.)

To illustrate the programmed editing of a connection table, again consider the table in Figure 2. Note that the connection between atom number 1 (the nitrogen atom) and atom number 2 (a carbon atom) is indicated in the table twice—once at rank #1 and once at rank #2 of the table. Similarly, all other attachments are recorded twice. Also note that for each noncarbon atom, the table explicitly indicates the number of attached hydrogen atoms. By using this built-in redundancy, the computer is able to detect most of the errors of transcription or omission introduced in keyboarding the table for computer input. When such an error is detected, the table is barred from further Registry processing and is instead filed for recycle and correction.

When a table successfully completes and passes the redundancy checks, it proceeds to a valence check. The program calculates the "acting valence" of each atom in the table and compares this valence to a stored list of the "common valences" for that element. If the acting valence is not on the stored list, then an error has been detected and the table is recycled for correction. As a supplement to this check, "uncommon" valences are always recorded in the input connection table so as to explicitly override the valence calculation checks made by the computer.

Once the valence checks have been successfully completed, the program checks the molecular formula. First, the program calculates a molecular formula from the information given in the connection table. This formula, including the hydrogen count, is compared to that calculated by the chemist and input by the keyboard operator. Any discrepancy between the two molecular formulas constitutes an error which causes the table to be recycled for correction.

In recycling a table for correction, the computer program identifies the type of error and prints a notice identifying the error. Some of the error notices which the program produces when an error is detected are listed in Table I.

As a further aid to error correction, the program in most cases stores and prints the erroneous connection table for recycled compounds. Examining the printed table enables the clerk to easily identify the error, which can then be corrected by rekeyboarding only that portion of the table found to be in error, and not the entire table.

When a correction is made to a table previously rejected, the entire checking process is repeated to be certain no new errors have been introduced and that all errors have, in fact, been corrected.

Table II gives a breakdown of some of the common errors detected by the computer programs and the relative frequency of error. As indicated in the table, we have found that 98% of the errors detected by the computer are introduced in the keyboarding operation.

ERROR SAMPLING TECHNIQUE

Although we instituted the computer error-detecting programs in the belief that they were sufficient to detect all but a vanishingly small number of the errors, the problem remained to verify this belief and to determine how many undetected errors were being admitted to the system. To audit this error level, we have established routine sampling and reprocessing procedures. Each week 100 or so previously processed structures are selected at random and submitted for reprocessing. This reprocessing presently begins with the structure and molecular formula as originally prepared by the chemist. In the future, however, CAS plans to carry the error study back to the structuring process so that we can determine the frequency with which errors are made in preparing structures.

For the structures selected for reprocessing, the entire registration process is repeated. To eliminate bias, reprocessing is done in such a way that to the clerical personnel and operators of the system, the entry is just another item of production. Only the Registry group supervisor knows which entries are part of the error study sample. When processing of the sample is complete, the source documents are retrieved from the processed material and the second registration is compared to the first. If both registrations result in the same Registry Number for the compound, then it is safe to assume that both were error free. If, however, the two registrations result in the assign-

Figure 2. An example Registry Form containing a chemist's hand-drawn structure and molecular formula and a clerically derived connection table.

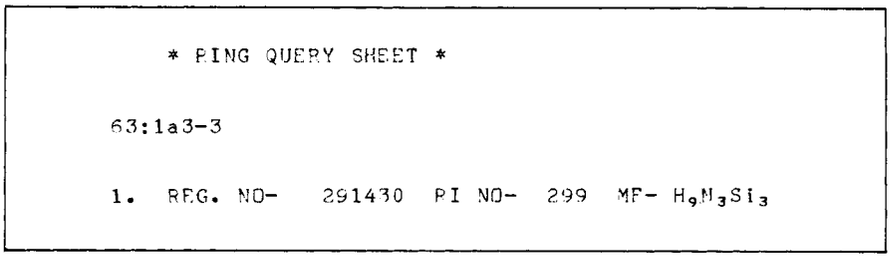

Table I. Error Notices for Recycled Connection Tables

Notices	Interpretation
Element Symbol Incorrect in Molform	An element symbol in the molecular formula is not valid according to the element table.
Molform Is Incorrect	The calculated molecular formula is not equal to the given molecular formula.
Hydrogen Count Is Incorrect	The hydrogen count in the molform does not equal the calculated hydrogen count.
Attachment Present but No Bond	Bond column blank but attachment is present.
Element Symbol Incorrect	Rank element symbol is invalid.
Acting Valence Incorrect	The acting valence of a noncarbon atom is not equal to the table valence.

ment of different Registry Numbers, then it is obvious that at least one was in error. When such a discrepancy is detected, the group supervisor traces back through the processing to find which process failed, and to identify the cause of the failure.

Although our main interest is in auditing the level of undetected errors and in determining the causes of these errors, the errors detected in this way are, of course, corrected. However, the most important results of the error study are that: (1) we are able to define the probable number of errors in the file, (2) we are able to identify the most error-prone operations, and (3) we are able to define the types of errors which go undetected. By knowing the level of error in the file, we can better judge the effectiveness of the system. By knowing which operations are most error prone, we are better able to adjust training programs and supervision practices in these areas. And by knowing the cause of undetected errors, we can more profitably allocate our manpower to detecting such errors and devising methods to reduce the chances that

Table II. Computer-Detected Errors in 33,257 Input Connection Tables

Transactions	Errors	% Clerical	% Chemist
Attachment Missing	1999	25.3	...
Attachment Incorrect	2021	25.5	...
Bond Incorrect	468	5.9	...
Molecular Formula Error	928	11.5	0.2
Element Symbol Missing	192	2.4	...
Hydrogen Missing or Incorrect Count	593	6.9	0.7
Other	1622	20.5	1.1
Total	7823	98.0	2.0

Table III. Error Study Results

Sample Size	4009
Number of Errors	20
Percentage of Errors	0.499
95% Confidence Interval	0.77

they will be made. Table III shows the results of the error study sampling to date. It shows that the indicated percentage of undetected errors is about 0.5%; the statisticians interpret this to mean that there is 95% confidence that the error level is no higher then 0.77%.

It is interesting to note that of the errors detected by the computer, over 98% were attributed to the clerical operations; however, of the errors admitted to the system undetected, nearly 30% were traced directly back to the chemists' handling of the stereochemistry (Table IV). To date no computer methods have been established which adequately identify errors in the stereochemistry recorded by the chemist and input by the keyboard operator.

From a purist point of view, an error is an error. However, from a more practical point of view, the six errors assigned to the chemists' handling of stereochemistry are partially counterbalanced by a feature of the Registry System that provides automatic cross references between stereoisomers, salts and their parents, and so forth. Therefore, all six of the erroneous registrations are nevertheless automatically cross referenced to the correct three-dimensional representation.

The next error shown on Table IV resulted from a program error. When this error was uncovered, we immediately corrected the program and identified the condition under which the edit program failed. We then located, examined, and corrected where necessary all compounds on file which met this condition and which might have also been admitted to the system in error.

The final major category of errors was that assigned to the clerical operations. These are divided into two subcategories: misinterpretation of structuring corrections and failure to cite the isotopic mass for a labeled compound. To reduce errors in the first category, we have

Table IV. Cause of Undetected Errors in Error Study

Cause	No. of Errors
Chemist Handling of Stereochemical Features	6
Error in the Computer Program	1
Clerical Input and/or Handling	
Misinterpretation of Structuring Conventions	10
Failure to Cite the Mass in the Table for isotopically labeled compound	3
Total	20

either provided additional training to the clerks or have stopped using certain unusual or ambiguous structuring conventions. To reduce errors in the second category, we have decided to introduce further redundancy by explicitly indicating labeled atoms in the molecular formula. As soon as this additional check is programmed, the chemists will indicate the presence of a labeled atom in the molecular formula as well as the structure. In this way the keyboard operator would have to make two errors of omission, which we feel is very much less probable than one error.

To summarize: The sampling procedures have taught us that significant reductions in the undetected error level can be achieved by devoting more of our effort to the handling of stereochemistry. Likewise the study has indicated a pattern in the clerical errors. With such information, we are better able to direct ourselves to the areas in which we can eliminate the most errors with the least expenditure of manpower and computer time.

COMPUTER-BASED ERROR DETECTION FOR REFERENCES AND NOMENCLATURE

Since we have now been able to reasonably assure ourselves that our handling of structures is adequate, we have renewed our emphasis on computer-based error control in the handling of bibliographic references and chemical nomenclature.

Presently we rely exclusively on clerical proofing to detect errors in the keyboarding of bibliographic references. We have, however, laid plans and established techniques to build into the *CA* references a form of redundancy in the form of a computer-generated check letter added to the *CA* abstract reference. This procedure will be implemented beginning with *CA* Volume 66. The check letter will serve as a means of detecting the common errors of transcription and simple alternations in the reference made during the keyboarding or manual handling. This detection will be made by computer program at the time of input. When an error is detected, the reference will be flagged and recycled so that the proofreaders can correct the input record before it is added to the permanent files.

The problems of identifying nomenclature errors by computer are, of course, quite different from those of identifying errors in structures. At present, the computer nomenclature error checks involve the analysis of conventions for punctuation, case, and font. For example, computer programs now automatically analyze nomenclature and insert correct italics and capitalization. Errors in italics and capitals are, however, errors primarily in form, not meaning, since the name can be understood without proper italics and capitals.

For most other errors in nomenclature, we rely heavily at present on the clerical proofing of computer-produced data sheets against the original source documents. However, CAS systems analysts and chemists are writing an algorithm for computer analysis of systematic nomenclature and automatic cross checking of the name with the connection table for correspondence of locants, prefixes, etc. The first step in solving this problem is to determine what types of nomenclature errors are being admitted to the system. This will be accomplished by a sampling procedure similar in concept to that used successfully in the structure work. We will select, at regular intervals, a sample of structures from the file and then retrieve from the computer all names for each structure. Chemists will then review the names and the structures to determine the accuracy of the name and the correspondence between name and structure. Based on the results of this study we hope to identify the most serious problems and begin concentrated effort aimed at solving them.

Until we can report automatic computer identification of errors in nomenclature, our most important tool in assuring the accuracy of nomenclature is constant use of the system and its data files in the registration operations.

As described by D. J. Whittingham *et al.* of the CAS staff, current registration operations involve the continual retrieval of nomenclature from the CAS files for examination by CAS chemists (2).

This retrieval and review provides additional checking and correction cycles to improve the accuracy of the Registry Files. Thus, the more CAS operations make use of the Registry, the more likelihood there will be that the errors in the Registry will be brought to light and corrected. This correction process in turn will make the Registry yet more valuable not only to CAS, but also to outside users. In sum, as in nearly any system, use becomes the means of and provides the impetus for the elimination of error.

LITERATURE CITED

(1) Leiter, D. P., Jr., Morgan, H. L., Stobaugh, R. E., *J. Chem. Doc.* **5**, 238 (1965).

(2) Whittingham, D. J., Wetsel, F. R., Morgan, H. L., "A Computer-Based Subject Index Support System," *ibid.*, **6**, 230 (1966).