

The Computer-Based Subject Index Support System at Chemical Abstracts Service*

D. J. WHITTINGHAM, F. R. WETSEL, and H. L. MORGAN**
Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received October 3, 1966

This paper describes a computer-based input system which reduces or eliminates many repetitive operations. This system reduces and conserves the over-all human effort required for input of structural, nomenclature, and bibliographic data while simultaneously improving the efficiency of the registration operation and increasing the reliability of the stored data.

Since early 1965, Chemical Abstracts Service has been developing an experimental computer-based Chemical Compound Registry System which is being supported by the National Science Foundation (NSF), the Department of Defense, the Food and Drug Administration, the National Institutes of Health, and the National Library of Medicine through a contract with NSF.

The data processed for this computer-based system involve structures, names, and references for compounds, primarily those processed in the current volumes of *Chemical Abstracts* (CA). Thus, each six months, the Chemical Compound Registry System receives data on about 200,000 compounds which are processed at CAS for *Chemical Abstracts*, and the *Ring Index* and its Supplements. Of these 200,000 compounds, about 163,000 have been reported previously in the literature, and have thus been processed previously at CAS. The remaining 37,000 compounds are newly reported chemical entities.

Without the computer-based support to be described in this paper, the registration of previously reported compounds for which new data continue to be documented in the literature would involve many highly repetitive operations—drawing of structures, calculation of molecular formulas, identification of ring systems, the derivation of systematic nomenclature, keyboarding, and editing, to name a few.

This input system is an integral part of the experimental CAS Chemical Compound Registry System which has been the topic of several recent papers (1), and will not be discussed in detail here. However, a brief review of the component parts of the Registry System is in order to clarify its relationship to the computer-based input system. The principal components of the Chemical Compound Registry System are three major computer files of compound-oriented data. The connecting link for these

three files is the permanent computer address for each compound—the Registry Number. The three files are:

1. The Structure File, which contains unique descriptions of the structural formulas including stereochemistry and isotopic labeling. The input to this file is subjected to an elaborate computer editing program, described by Leiter and Morgan (2).
2. The Nomenclature File, which contains all of the names for a given compound available in a variety of sources. These names are coded as to type—e.g., preferred CA index name and trade name. Presently, editing for the Nomenclature File is done mainly by chemists during the production of the CA indexes. The existence of the Nomenclature File makes possible the printing of a variety of special indexes—e.g., an index of trade names versus the CA preferred index names.
3. The Bibliography File, which contains references to the CA abstracts. By use of other existing computer files these references can be coordinated with the corresponding original journal references.

These files of the experimental CAS Chemical Compound Registry System are reviewed for accuracy through the efforts of CAS chemists and clerical personnel, aided by the computer-based input system described in the following pages. Two types of computer support are provided through this system, one based on the Structure File, and one based on the Nomenclature File.

SUPPORT THROUGH THE STRUCTURE FILE

The first set of procedures, based on the Registry System Structure File, is applied to compounds registered from those sections of CA reporting the highest percentage of new compounds—namely, the synthetic organic sections.

Compounds selected for registration from these sections have their structures drawn and molecular formulas cal-

* 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.

culated by CAS professional staff. The structural formulas are then clerically processed for registration. In this operation, all nonhydrogen atoms in the structural formulas are numbered in any convenient sequence. Then the atoms and bonds are keyboarded in tabular form for input to the computer. In the same operation the CA reference, any trivial names, and the calculated molecular formula are keyboarded for computer input (3). The structure is then registered, a process in which a computer program compares the structural information input with the data on file for all other structures. Registration results in one of two situations.

The first, a "hit," means that the compound has been registered previously and thus is already on file. In such cases, the computer program retrieves from the Registry Files the molecular formula, the edited and correctly formatted index name(s), and the Registry Number for the compound. This information, together with the originally keyboarded CA reference for the current volume, is printed by the computer on a data sheet (Figure 1) and sent to a CA chemist for review. All of the information on this sheet is reviewed for accuracy and corrected for discrepancy as necessary.

Through the information retrieved from the Registry File, the input system has reduced the effort required for registration. That is, compounds for which hits result do not require naming, and the names do not require keyboarding, editing, or re-entry into the computer.

Additionally, if the compound contains a ring system, the computer programs will determine whether the ring system is new or has been registered previously. (Records for some 18,000 ring systems are on file as a result of the registration of the *Ring Index* and its Supplements.) Where there is a hit, the computer prints a data sheet (Figure 2) giving the Registry Number, Ring Index Number, and molecular formula of the ring system. This saves the effort of renaming and reregistering the ring system.

The second situation that may obtain as a result of registration is "no hit," meaning that the compound is new to the Registry System. In such cases, the computer prints a data sheet (Figure 3) containing the molecular

formula and the CA reference for the current volume (that is, the data that were input to the computer for registration). These data are proofread to assure that they have entered the computer accurately, and then CAS professional staff derive the CA index name(s), which are keyboarded and entered into the Registry Files.

SUPPORT THROUGH THE NOMENCLATURE FILE

The set of input procedures based on the Registry System Name File applies to sections of CA other than the synthetic organic sections. These sections contain a high percentage of compounds reported previously in the literature.

CAS staff concerned with the selection of compounds to be registered dictate the available systematic or trivial names of the selected compounds and the corresponding CA references. This information is then keyboarded and input to the computer, where each name is compared with the names already filed in the Nomenclature File. As with Structure File support, one of two situations obtains when a systematic or trivial name is compared with names on file in the computer.

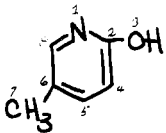
The first, a "hit," means that the input name is already on file and therefore that the compound has been registered previously. In these cases, the computer program retrieves from the files the molecular formula, the edited and correctly formatted CA index name(s), and the Registry Number for the compound. This information is printed by the computer together with the name and CA reference dictated at the time of the compound selection. Figure 4 illustrates such a computer-printed proof sheet.

Through this procedure, the computer input system has again reduced the effort required for registration. For compounds for which hits result, the structural formulas do not have to be drawn, the molecular formulas do not have to be calculated, and the preferred CA index names do not have to be generated, keyboarded, edited, or re-entered into the computer.

The second situation that may obtain after name input is "no hit," meaning that the name is not on file. For

40764	DATA CHANGED	SATURDAY, JUNE 18, 1966
	TO INDEXING	
Vol 64	Sec 42- 7	Start 9783e 1 End 9791a 0 Ind XXX Typ DRR Dat-66167
64:9789f2-3		
F	MF *	C ₁₅ H ₂₆ O
R	PINH *	Androsta-3,5-dien-17-one
C	ID	64:9789f2-3
	T/R	1912636

Figure 1. Structure file "hit" computer proof sheet.

63 Vol.		Sheet No. 33		Reg. No.	
34-1 Sec. No.		Chem. Date			
Page 51493		Compd. No. 7		Codes	
Prop.		Abst. Des. Auth. Name 2			
TID N° 135086 K		File		MF C ₆ H ₇ NO	
					
Stereo NS Stereo Code					

At No.	Elem.	Group	B D	Att	B D	Att	B D	Att	B D	Att	B D	Att	B D	Att	B D	Att	CH	ABV	ABM	Hyg
1	N		2	8	1	2														
2	C		2	4	1	2	1	3												
3	O		1	2																1
4	C		2	2	1	5														
5	C		2	6	1	4														
6	C		1	8	2	5	1	7												
7	C		1	6																
8	C		2	1	1	6														
9																				
10																				
11																				
12																				
13																				
14																				
15																				
16																				
17																				
18																				
19																				
20																				

Figure 2. Computer-produced ring query sheet.

such situations, structural formulas must be drawn and the molecular formulas calculated. The registration process then proceeds, with further support obtained as previously described under Support Through the Structure File.

Even in the "no hit" situation, however, the name and associated data have been added to the computer record. The next time the same name is encountered in CA indexing, computer support through the Name File will be possible.

SOME SAVINGS OF CLERICAL EFFORT

Two areas where the computer-based support system has effected considerable savings in the use of clerical

effort involve the single keyboarding of preferred index names and CA references, and the use of keyboarding "shortcuts" that save keystrokes during data input.

CA index operations currently use as manuscript 3 × 5 cards containing one entry per card. For compound entries, the name of the compound and the CA reference appear on two cards, one for the Subject Index, and one for the Formula Index. Before the computer-based input system was developed, these 3 × 5 cards were typed directly. In addition the registration operation then required another keyboarding of many compound names and CA references. These operations are now all combined into a single keyboarding of data for entry into the computer, which then delivers the index cards and records the appropriate information in the Registry. The total

COMPUTER-BASED SUBJECT INDEX SUPPORT SYSTEM AT CAS

40765	DATA CHANGED	SATURDAY, JUNE 18, 1966
	TO INDEXING	
Vol 64	Sec 42- 7	Start 9783e 1 End 9791a 0 Ind XXX Typ DRR Dat-66167
64:9789f2-4		
F	MF *	C ₂₅ H ₂₈ N ₂ O ₅
C	ID	64:9789f2-4
	T/R	4975466

Figure 3. Structure file "no hit" computer proof sheet.

keyboarding effort has been reduced by an estimated 25% as a result of this procedure.

The use of keyboarding shortcuts has also been made possible through the use of the computer input system. Under the old system, the typists were never afforded the possibility of shortcuts. For example, the registration of six different esters of 2,4,6-triiodobenzoic acid required the keyboarding of the parent acid name six times.

In the computer-based input system, however, the typists now keyboard the complete name of the parent acid once. For the five remaining esters, a two-character "ditto code" is typed instead of the parent acid name (Table I). The ditto code instructs the computer to print out the parent acid name keyboarded for the first entry. Note that the ditto code is fully expanded by the computer and thus is not carried in the permanent files. Therefore, it makes no difference in the stored data whether or not the typist uses the shortcut.

Two other ditto features are also used. One is used to repeat a name's modification (that portion of the name that appears in light-face type in the CA Subject Index) from one entry to another. The second, used in registering alphabetized lists of names, repeats that portion of a

name up to and including the first comma followed by a space—*e.g.*, the comma of inversion. We estimate that 42,000 keystrokes are saved by these three features every work day; this is equivalent to 5% of the total keyboarding effort.

PROJECTED SUPPORT IN PRINTED ISSUE, VOLUME, AND COLLECTIVE INDEXES

In 1967, CAS expects to install a much more advanced computer support system, which will be directed primarily at support of index operations rather than Registry operations. Through this system, CAS will eliminate the use of the 3 × 5 manuscript cards, and, instead, produce "camera-ready" copy from the computer for final proof and printing operations. This will result in greater computer support for the indexing operations. For example:

1. Alphabetizing of the entries for an index will be performed as a computer operation.
2. The merging of edited index volumes to collective indexes will also be performed as a computer operation.

33156	NEW WORKSHEET	TUESDAY, JULY 26, 1966
	NAME MATCH PERFORMED	
Vol 64	Sec 67- 7	Start 10220e End 10233b 0 Ind DL Typ LCS Dat-66200
64:10223d1-2		
F	MF *	C ₆ H ₁₃ NO ₂
R	PINH *	Hexanoic acid, 6-amino-
K ₆	EAINH *	ε-Aminocaproic acid
C	ID *	64:10223d1-2
	T/R *	60322

Figure 4. Name file "hit" computer proof sheet.

Table I. Example Use of "Ditto Code" to Save Keystrokes

Typed as	Computer-printed as
Benzoic acid, 2,4,6-triiodo-, ethyl ester	Benzoic acid, 2,4,6-triiodo-, ethyl ester
Ipmethyl ester	Benzoic acid, 2,4,6-triiodo-, methyl ester
Ippropyl ester	Benzoic acid, 2,4,6-triiodo-, propyl ester
Ipisopropyl ester	Benzoic acid, 2,4,6-triiodo-, isopropyl ester
Ipbutyl ester	Benzoic acid, 2,4,6-triiodo-, butyl ester
Ip3-nitropropyl ester	Benzoic acid, 2,4,6-triiodo-, 3-nitropropyl ester

Further, new indexes and a data base for searching property, reaction, and use information from the computer-based Subject Index Support System will become a reality.

Our users will receive the first products of this completely mechanized system in 1967 in the form of the computer-composed volume Author and Formula Indexes. Note, however, that the Registry Support System described in this paper is in operation at CAS at this time.

APPENDIX

Explanation of Terms Used on Computer-Produced Data Sheets

The data sheets produced by the computer as described in the foregoing text carry several types of information for use in entering compounds into the CAS Chemical Compound Registry System. The sheets are printed in worksets grouped by CA sections and by column fractions, and within a given workset, each compound is represented by a single data sheet. The following is an explanation of the terms used on the data sheets in Figures 1, 3, and 4 of this paper.

The heading of each sheet includes a sequential number assigned by the computer, an identification of the type of sheet—e.g., "new worksheet"—and the day and date on which the data were processed by computer.

The following data are keyboarded once for each workset and are then automatically printed by the computer on each applicable data sheet.

Vol	The volume of CA from which the data were obtained.
Sec	The section and issue number of CA from which the data were obtained.
Start-End	The limits (expressed as column numbers, letter fractions, and abstract numbers of the CA issue) between which the data were obtained.
Ind	The chemist, identified by initials, who dictated the data. (The examples use XXX as the indexer's initials.)

Typ The typist, identified by initials, who keyboarded the data.

Dat The date the data were keyboarded.

The following are codes for major fields keyboarded for each data set:

F	Formula.
R	Preferred CA index name.
N	Added CA index name.
K ₁₋₅	Extra added CA index name.
K ₆₋₉	Fields in which systematic, trade, or trivial names are input for Index Support through the Nomenclature File.
C	Identification or reference.

The following are codes for subfields printed automatically by the computer:

MF	Molecular formula.
PINH	Preferred index name heading, that portion of the index entry that appears in bold-face type in the subject indexes.
EAINH	Extra added index name heading.
ID	The CA reference including column number, letter fraction, abstract number, and compound number. The latter two numbers are for internal computer use only.
T/R	The temporary identification number (T) or the Registry Number (R) of the compound. Temporary identification numbers are used in initial processing until a Registry Number is assigned. The Registry Number becomes a compound's permanent identification in the CAS Chemical Compound Registry System.

LITERATURE CITED

- (1) See, for example, Leiter, D. P., Jr., Morgan, H. L., Stobaugh, R. E., *J. Chem. Doc.* **5**, 238 (1965).
- (2) Leiter, D. P., Jr., Morgan, H. L., "Quality Control and Auditing Procedures in the Chemical Abstracts Service Compound Registry," *ibid.*, **6**, 226 (1966).
- (3) These clerical operations are described more fully in Leiter *et al.*, *ibid.*, **5**, 240-1 (1965).