ARTICLES

Chemical Abstracts Service Chemical Registry System. 10. Registration of Substances from Pre-1965 Indexes of *Chemical Abstracts*

KAREN A. HAMILL, R. DAVID NELSON, GERALD G. VANDER STOUW, and ROBERT E. STOBAUGH*

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

Received February 24, 1988

The Chemical Abstracts Service Chemical Registry System, operating since 1965, uniquely identifies chemical substances on the basis of molecular structure. Chemical Abstracts Service is now registering chemical substances cited in indexes to *Chemical Abstracts* prior to 1965. This effort will result in several hundred thousand additional chemical structures, along with their names, being available for online searching in the Registry File. Both the newly registered substances and those already on file are being linked to their pre-1965 citations in *Chemical Abstracts* in a new file called CAOLD. In this effort the printed Formula Index entries are converted to computer-readable form by using optical character recognition with the data subsequently processed with existing computer programs.

INTRODUCTION

The Chemical Abstracts Service (CAS) Chemical Registry System, which became operational in 1965, identifies chemical substances uniquely on the basis of molecular structure. A computer-generated structural representation called a connection table is the means for recording substances. The design, operation, and capabilities of the Registry System are detailed in a series of papers by Stobaugh et al.[1-9] To date, over 8 million substances have been added to the Registry File, which is available for search and display via STN International.[10]

Beginning in 1965, substances appearing in the *Chemical Abstracts* indexes were routinely added to the Registry System. CA citations for these substances from 1967 forward have been placed in the CA File, also available via STN International. It has long been recognized that the chemical substances cited in the indexes to *Chemical Abstracts* prior to 1965 were also needed for online search and display. These substances would not be on file unless they have appeared in the *Chemical Abstracts* indexes since 1965 or were registered as part of a special collection, such as a chemical handbook.

The changing and expanding role of the CAS Chemical Registry System, progressing from an in-house processing tool to a resource recognized worldwide as an authority for chemical substance identification, emphasized the importance of registration of pre-1965 chemical substances. Such an augmentation of the Registry File would allow researchers to apply the unique advantages of online, interactive structure searching to the retrieval of substance information documented decades ago, information found only in printed volumes and extracted by manual effort. Benefits of registration from older indexes include the following: replacing printed index searching with online searching; saving time in patent and novel-substance searching; and searching by structure rather than names for classes of compounds.

An additional stimulus for the registration of chemical substances indexed in *Chemical Abstracts* before 1965 was provided by the emergence of computer technologies for handling the task. Optical character recognition, which converts the printing on paper into a computer-readable form, had been developed to the point where speed and accuracy were adequate for this application. Computer programs to convert chemical names into chemical structural representations had already been developed by CAS; these programs could be expected to considerably reduce the amount of manual effort needed to process names once they were in computer-readable form.

A successful feasibility study led to a funding drive to get the work under way. A production operation was established to convert the printed indexes into computer-readable form followed by conversion of chemical names into structures.

THE FEASIBILITY STUDY

In its 1978 report to the United States Congress pursuant to Section 25(b) of the Toxic Substance Control Act (PL 94-469), the Council on Environmental Quality recommended that U.S. government agencies use Chemical Abstracts Service Registry Numbers as the standard for identification of chemicals and chemical mixtures in their files. As a result of this recommendation, the Council contracted with Chemical Abstracts Service to study the technical feasibility and resource requirements for registering chemical substances indexed in *Chemical Abstracts* before 1965. This study would identify the kinds of *Chemical Abstracts* indexes—Subject Index, Formula Index, or both—from which the chemical substance names were to be extracted, the number of chemical substances expected to be registered, methods for converting printed index pages into computer-readable form, registration methods available, and the estimated costs.

Analysis of the *Chemical Abstracts* Formula Indexes. The feasibility study found that the *Chemical Abstracts* Formula Index, rather than the Subject Index, would be the better source of data. Both indexes contain the names of substances, but the Formula Index contains the molecular formula, an important tool in editing substance data during registration. Additionally, in the Subject Index the substance names are intermixed with data extraneous to registration, and these data cannot easily be removed. However, there would be situations, as discussed below, where it would be necessary to look up specific information in the Subject Index.

0095-2338/88/1628-0175$01.50/0

Table I. Collective Indexing Periods for *Chemical Abstracts* Formula Indexes and Estimated Number of Chemical Substances To Be Processed

index	years covered	no. of substance entries
7th Collective Formula Index (pre-1965 portion)	1962–1964	578 000
6th Collective Formula Index	1957–1961	510 000
5th Decennial Formula Index	1947–1956	463 000
27-Year Collective Formula Index	1920–1946	328 000
total		1 879 000

Processing of the pre-1965 Formula Index data would be in reverse chronological order at a rate determined by available staff and financial resources. Rather than processing the semiannual volume indexes, the collective indexes would be used. These were, and still are, prepared by merging several years' worth of *Chemical Abstracts* Formula Indexes together at intervals to reduce the repetitive work that would otherwise be needed for a volume-by-volume search. When this merging was done, substance names were "brought up to date" to reflect the current nomenclature rules for improved consistency. These collective indexes cover a 45-year span and are listed in Table I. The estimated number of chemical substances to be registered for each collective period is also given. No *Chemical Abstracts* Formula Indexes were prepared prior to 1920.

Registration of the chemical substances in the 1920–1964 period would present some unique problems for which solutions would be needed. For example, in some cases, only line formulas were given in the primary document; this information was carried over into the *Chemical Abstracts* indexes. However, since more detail is needed to generate a name or chemical structure, these entries would not be processed. Because of modifications in Chemical Abstracts Service editorial policies, which generally reflect the changes in and growth of chemistry and chemical engineering, indexes produced during various collective periods are not all alike. Prior to 1962 all positional isomers of a given chemical substance were posted to a single heading in the Formula Indexes. This was not the case for the corresponding Subject Indexes, where the specific isomers were listed whenever possible. Following is an example of an entry from the *Sixth Collective* Formula Index and several more specific entries from the Subject Index to which it corresponds.

Sixth Collective Formula Index
$C_{16}H_{14}$
Anthracene, dimethyl-, 51:324i, 340g, 353b, 870g, 9330a, 10241c, 11080g, 13580b, 14664c; 52:2827g, 5282a, 6381i, 7254f, 11800a, 16253g, 16843e, 16917e, 19394f; 53:2782g, 2881i, 6768h, 18458d; 54:4154h, 5424a, 8277e, 9857g, 17045e, 18066g, 19616d;

Sixth Collective Subject Index
Anthracene, dimethyl-, isomers, detection in gas oil, 53:18458d
---, 1,4-dimethyl-, 52:7254f
---, 2,3-dimethyl-, plastic scintillators contg., 52:16917e
---, 2,6(or 2,7)-dimethyl-, spectrum of, 51:10241c
---, 2,10-dimethyl-, and picrate, 52:6381i

Approximately 25–30% of the entries in the collective Formula Indexes to be examined had isomers grouped in this way. Since identification of individual isomers using the corresponding Subject Indexes would have required considerable manual effort, that step was not undertaken in the feasibility study. It was recognized that the step should be included in actual production, and, as noted later, this proved to be the case.

Table II. Estimated Time for Keyboarding and Proofing Collective Formula Indexes

index time period	keyboarding time, h	proofing time, h
1962–1964	8500	4200
1957–1961	7600	3800
1947–1956	8400	4200
1920–1946	3900	1900

The feasibility study also examined the characteristics of the collective Formula Indexes to detect areas where processing problems may occur. As is typical of printed indexes, the Formula Indexes did not display duplicated portions of multiple, consecutive entries.

Sixth Collective Formula Index
$C_{16}H_{14}$
Fluorene, 9-allyl-, 55:491d, 17594g
---, 9-isopropylidene-, 52:10033a, 53:4908e, ...
---, 9-methyl-9-vinyl-, 51:3535h

It would be necessary to convert such merged index entries into complete, stand-alone "records" for subsequent processing steps. Each record would consist of the molecular formula, the full name of the chemical substance, and the associated *Chemical Abstracts* reference(s). To prepare such records for later parts of the feasibility study, an experimental Name Record Builder program was written; the following are examples of its output from the records shown above.

1. Molecular Formula: $C_{16}H_{14}$
   Name: Fluorene, 9-allyl-
   Reference: 55:491d, 55:17594g
2. Molecular Formula: $C_{16}H_{14}$
   Name: Fluorene, 9-isopropylidene-
   Reference: 52:10033a, 53:4908e, ...
3. Molecular Formula: $C_{16}H_{14}$
   Name: Fluorene, 9-methyl-9-vinyl-
   Reference: 51:3535h

This line-by-line format produced by the Name Record Builder allowed for easy and rapid proofing. Comparisons could be made with the original page copy instead of having to compare lines of different data content. Procedures for recycling corrections of the input data were not investigated since, during production, proofing would be online rather than by use of a printed copy.

Conversion to Computer-Readable Form. Before records from the pre-1965 indexes could be processed, it would be necessary to convert them from printed to computer-readable form. There were two methods for doing this conversion: keyboarding and optical character recognition. The feasibility study evaluated both methods for their reliability and economy as methods for this conversion.

Table II summarizes the estimated time required to key and proof the data in the collective Formula Indexes. The estimates, which examined pages selected from indexes spanning the 45-year period, are based on copying the material line by line from each page along with appropriate processing identification codes. Proofing of references would be required since only limited computer edits might be applied. According to general experience at Chemical Abstracts Service, the time required for proofing using printed proofs is approximately half the keyboarding time.

The keyboarding–proofing effort was compared to optical character recognition. At the onset of this work, several systems to translate printed characters into computer-readable form by optical character recognition were on the market. The

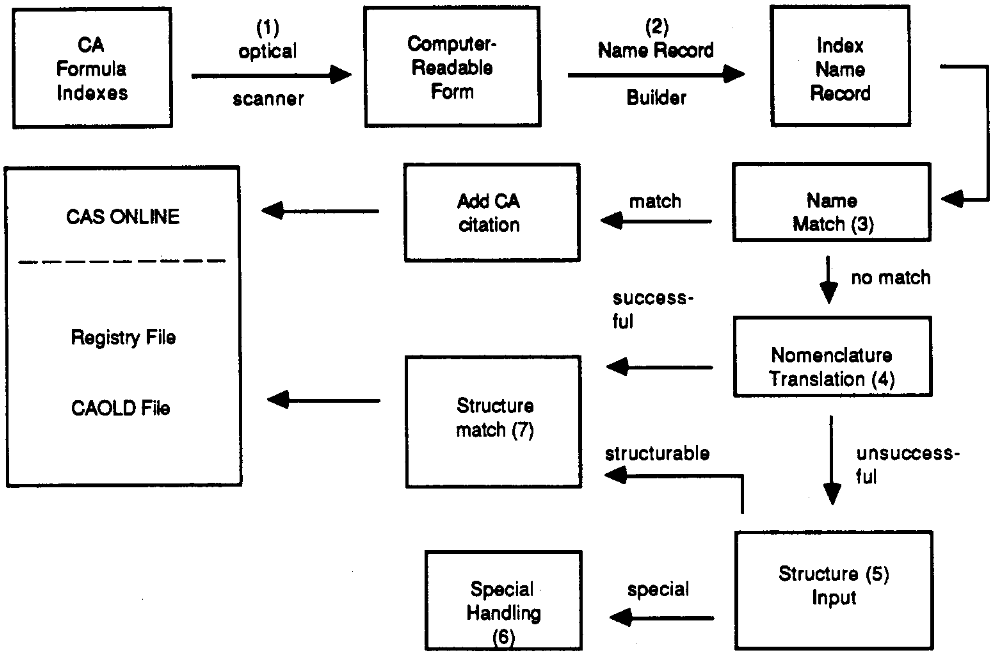

Figure 1.

Kurzweil data entry machine, produced by Xerox Corp. subsidiary Kurzweil Computer Products, Inc., was selected because of its ability to recognize virtually any typeset material.[11]

Conversion tests of sample pages from the Formula Index were conducted. The Kurzweil data entry machine at first had considerable difficulty handling the small size and juxtaposition of characters within a line as well as the close proximity of the lines. However, once the print on these sample pages was enlarged, by use of a copying machine, the ability of the machine to read the data accurately increased dramatically. The Kurzweil data entry system has a cathode-ray terminal that displays lines of text as they are scanned. Those characters whose interpretation is questioned by the machine are highlighted and can be quickly resolved at the time of scanning by the operator of the terminal. The tests showed[11] that only 0.31% of the characters in the computer-readable version of the printed index were in error. This very high accuracy suggested that no proofing would be required for the data resulting from scanning, especially if the scanning were combined with computer edits for identification of inconsistencies and detection of errors in the data. Subsequent experiments, including a model of a production environment, showed this to be the case.[12]

Comparison of cost estimates for the keyboarding–proofing procedure versus the use of optical character recognition, including the cost of acquiring the Kurzweil data entry machine, revealed the latter to be more economical.

Registration Procedures. After the names from the sample pages of printed Formula Index entries were converted into computer-readable form and subsequently processed into complete records by using the Name Record Builder, they were ready for automated registry processing. Three methods were available:

(1) Name match. An input name is matched with names already on file; the CAS Authority Database currently contains more than 12 million names for the more than 8 million substances on file.

(2) Nomenclature translation. Developed in the late 1960s,[13,14] nomenclature translation is the automatic conversion of a systematic chemical substance name into an atom-bond connection table.

(3) Structure input. A structure drawing is prepared by a chemist followed by keyboarding of that structure.

The feasibility study evaluated some possible orders in which these techniques might be applied and developed the workflow, which is shown in simplified form in Figure 1. This workflow includes the following steps:

(1) A Formula Index record is converted to computer-readable form by using the optical scanner.

(2) The name is then processed by the Name Record Builder program to prepare an individual name record.

(3) The name is processed by name match. If it matches a name on file, the existing Registry Number is retrieved. The pre-1965 CA citation for that Registry Number is added to CAS files (in practice, a new file called CAOLD was established for these references). If there is no match, the name is sent to nomenclature translation.

(4) The name is processed by nomenclature translation. If a connection table is generated, that table is sent to be matched against the structure file of the Registry System; if not, the name is sent to a chemist to review. (When the production operation was established, it was found useful to have names that failed nomenclature translation reviewed by a professional editor; those that were corrected by the editor were reprocessed by the sequence of name match and nomenclature translation.)

(5) The chemist prepares a structure diagram, which is keyboarded for entry to the Registry System and conversion to a connection table. Connection tables resulting either from structure input or for nomenclature translation are matched against the records already present in the structure file of the Registry System. If they match, the existing Registry Number is retrieved; if not, a new number is assigned. In either case, the CA citation is added to the CAOLD file.

(6) The feasibility study identified a variety of situations in which the information in the pre-1965 Formula Index was not complete enough to permit the above processes to be followed. These include, for example, the use of trivial or trade names that were not on file, so that the substances could be identified only by consulting original documents no longer available at CAS; such lookup was judged to not be economically feasible. There are also entries such as "Anhydride", "Compd.", and the like for which no structural information was available. The incidence of these situations increases as one goes farther back in time. There are also a number of entries in the Formula Index that are cross-referred to the Subject Index; i.e., the number of CA citations was so high that these citations were not included in the Formula Index.

Table III. Estimated Number of Chemical Substance Names and Their Methods of Registration

	expected registrations (all numbers in thousands)				
index time period	name match	nomenclature translation	structure input	"incomplete" records	records requiring lookup of specific isomers
1962–1964	158 (27.3%)	229 (39.6%)	149 (25.8%)	42 (7.3%)	0
1957–1961	105 (20.6%)	195 (38.2%)	110 (21.6%)	38 (7.4%)	62 (12.2%)
1947–1956	78 (16.8%)	182 (39.3%)	99 (21.4%)	44 (9.5%)	60 (13.0%)
1920–1946	41 (12.5%)	89 (27.1%)	70 (21.4%)	81 (24.7%)	47 (14.3%)

Table IV. Number of Chemical Substances and Their Methods of Registration for the *Seventh* and *Sixth Collective Indexes*

	registrations by (all numbers in thousands)			
index time period	name match	nomenclature translation	structure input	"incomplete" substances
1962–1966	390 (40.2%)	240 (24.7%)	327 (33.7%)	14 (1.4%)
1957–1961	161.5 (36.0%)	167.5 (37.4%)	116.0 (25.9%)	3.0 (0.7%)

These situations will normally be set aside, and no registration will take place.

(7) Structures added to the Registry System are added to the Registry File of CAS ONLINE; CA citations are added to the CAOLD file.

The above workflow was recommended by the feasibility study and became the basis for the workflow used in production operations when the actual processing was started. The eventual production workflow was, of course, somewhat more complex than this simplified scheme indicates. For example, some situations were encountered in practice in which the nomenclature translation programs gave incorrect translations that had not been encountered in program testing; edits were written and added to the flow to prevent these names from going to the translation program.

The feasibility study also prepared estimates of the numbers of records from each time period that would lead to registrations by the various methods described above. These estimates are given in Table III. The percentages in each case refer to the total number of records to be processed from the particular time period shown.

FUNDING FOR THE PRE-1965 REGISTRATION PROJECT AND PILOT TESTING

Once the feasibility study had demonstrated that it was technically feasible to register substances from the pre-1965 Formula Indexes, estimates were prepared for the costs of actually doing the job. It was estimated, in 1982, that going back all the way to 1920 would require at least $5 million and would require a minimum of four years elapsed time. It was then proposed to begin the job once funding of at least $2.4 million was assured, so that at least a substantial portion of the job could be done, and proceed back in time as resources would permit. Because the pre-1965 registration would require a substantial investment before any revenue could be realized as a result of it, CAS decided to seek funding from organizations who would be expected to use the resulting data. Following approval by the Board of Directors of the American Chemical Society at the end of 1982, a funding effort was undertaken, and the amount needed to begin the effort was raised by mid-1983. As a result of presentations before representatives of companies and organizations in the U.S. and Canada, along with direct mailings to European and Japanese firms, more than 100 organizations had contributed amounts ranging from $16 000 to over $250 000 by the end of 1983.[15] Thus, the registration project could be started and could proceed according to available funding.

The first step was a pilot test to firmly establish the requirements and determine the needed resources for the design, development, and implementation of a full production system. In this pilot, samples of names from the *Seventh* (1962–1966) and *Sixth* (1957–1961) *Collective* Formula Indexes were optically scanned. Programs were prepared for editing the computer-readable material resulting from scanning and converting it into production workunits. The samples were processed by using name match and nomenclature translation. Positional isomers of a given chemical substance posted to a single heading in the *Sixth Collective* Formula Index had to be individually identified by technical staff; the associated costs were recorded. Workunits added or changed were recycled through name match and nomenclature translation to see how many additional registrations were achieved. Finally, estimates were made for the costs of processing the data from the *Fifth Decennial* (1947–1956) and *27-Year Collective* (1920–1946) Formula Indexes, since page and line formats in these indexes differed from the more recent indexes. The pilot study was then used as the basis for the detailed workflow to be used in actual production operations.

PRODUCTION OPERATIONS

Regular production for the pre-1965 registration project began with the *Seventh Collective Index* data. The *Seventh Collective Index* period encompassed 1962–1966, so it included two years from which substances had already been registered. It was determined, however, that it would be more economical to process the total collective material rather than separate out only that part that required processing. In addition, coordination compounds and incompletely described substances, not covered previously, were included for registration.

Pages of the Formula Index were batched 5–10 at a time and scanned with the Kurzweil data entry machine, which was operated 16 h per day, 5 days per week. As each page was scanned, its image would appear on a cathode-ray tube connected to the scanner. Characters that the scanner had difficulty reading or could not read at all were highlighted at the terminal; the machine operator could resolve them. These data were then sent via direct communication line to the UNIX operating system, where several kinds of edits were applied: letters in the line formulas were checked for alphabetization; subscripts in excess of five digits were highlighted; peculiar letter combinations, usually resulting from scanning errors, were diagnosed; etc. This edited material was then forwarded for processing through name match and nomenclature translation.

If a chemical substance name matched against a name already in the authority database of chemical names, the CA citation was added to the CAOLD File associated with the retrieved Registry Number; an appropriate flag such as "7CI" or "6CI" corresponding to the source period was added to the name record. If the name did not match, it was sent to nomenclature translation. If it was successfully translated, the resulting connection table was sent to be matched against the structure file. If it could not be converted to a connection table, it was checked for errors. Any errors were corrected, and the name was again submitted to name match. If still no match occurred, the record was sent to nomenclature translation. If the name could not be translated, it was routed to a chemist who prepared a structural diagram for the compound. This diagram was converted to computer-readable form and once again compared to existing structures in the Registry File. If this last comparison did not retrieve an existing structure, a new Registry Number was assigned.

The CAOLD File has been available to users since March 1984 and continues to grow as the pre-1965 registration project continues. As of June 1988, the processing of the *Seventh Collective* Formula Index was complete and the *Sixth Collective* Formula Index was more than half done, with structure input being currently in progress. Table IV gives data on numbers and kinds of registration for these indexes. It must be noted that the decision to process the full *Seventh Collective Index* means that the data for 1962–1966 cannot be readily compared to the projections for 1962–1964 shown in Tables I and II; the total volume of data processed obviously is much higher, since material for five years is being processed rather than three years. Also, since two of the five years had already been registered, the percentage of material successfully name matched was much higher than the projections.

Processing of data from the *Seventh Collective Index* (1962–1966) has resulted in a total of 567 000 connection tables being matched against the Registry File (240 000 from nomenclature translation and 327 000 from structure input). Of these, 323 000 substances (182 000 from nomenclature translation and 141 000 from structure input) were new to the Registry System. Of all the records processed from the *Seventh Collective Index* (including those that resulted in successful name match), 634 000 (65.3%) matched records already on file.

SUMMARY

The sequence of operations for registering substances reported prior to 1965 in the *Chemical Abstracts Collective* Formula Indexes is (1) converting the printed data to computer-readable form by using optical character recognition, (2) editing to standardize format, (3) translating the data into discrete records consisting of molecular formula, index name, and reference(s), (4) identifying and isolating those records that contain substances that are cross-referred to the Subject Index or might only be unambiguously defined by using the source document, (5) using automated Registry processing to process the chemical substances in the order of priority, name match, nomenclature translation, structure input, and (6) routing "new" chemical structures to the Registry File and citations to both new and existing substances to the CAOLD File. Processing for the *Seventh Collective Index* (1962–1966) was completed and processing for the *Sixth Collective Index* (1957–1961) more than half completed as of June 1988. Current plans call for completion of processing of the *Sixth Collective Index* data. Any decision to proceed further back in time will depend on availability of funding.

ACKNOWLEDGMENT

The Council on Environmental Quality, in cooperation with the Environmental Protection Agency and the National Library of Medicine, has provided financial support for the feasibility study. This assistance is gratefully acknowledged. A number of companies have contributed to the cost of registration of substances from pre-1965 indexes. This assistance is gratefully acknowledged.

REFERENCES AND NOTES

(1) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Chemical Registry System. 1. General Design". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111–121.

(2) Freeland, R. G.; Funk, S. A.; O'Korn, L. J.; Wilson, G. A. "The Chemical Abstracts Service Chemical Registry System. 2. Augmented Connectivity Molecular Formula". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 94–97.

(3) Blackwood, J. E.; Elliott, P. M.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Chemical Registry System. 3. Stereochemistry". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 3–8.

(4) Vander Stouw, G. G.; Gustafson, C. R.; Rule, J. D.; Watson, C. E. "The Chemical Abstracts Service Chemical Registry System. 4. Use of the Registry System To Support the Preparation of Index Nomenclature". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 213–218.

(5) Zamora, A.; Dayton, D. L. "The Chemical Abstracts Service Chemical Registry System. 5. Structure Input and Editing". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 219–222.

(6) Stobaugh, R. E. "The Chemical Abstracts Service Chemical Registry System. 6. Substance-Related Statistics". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 76–82.

(7) Mockus, J.; Stobaugh, R. E. "The Chemical Abstracts Service Chemical Registry System. 7. Tautomerism and Alternating Bonds". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 18–22.

(8) Moosemiller, J. P.; Ryan, A. W.; Stobaugh, R. E. "The Chemical Abstracts Service Chemical Registry System. 8. Manual Registration". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 83–88.

(9) Ryan, A. W.; Stobaugh, R. E. "The Chemical Abstracts Service Chemical Registry System. 9. Input Structure Conventions". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 2–28.

(10) Rich, C. "Searching STN International". *Database* **1986**(Oct), 116–120.

(11) Nelson, R. D.; Hamill, K. A. "Optical Scanning at Chemical Abstracts Service for Building Computer Files from Printed Index Data". *Recognit. Technol. TODAY*, **1985**, *7*(5), 1–6, 15.

(12) Anonymous. "Computer Reads, Computer Speaks". *Online Database Rep.* **1980**, *1*(9), 69.

(13) Vander Stouw, G. G.; Naznitsky, I.; Rush, J. E. "Procedures for Converting Systematic Names of Organic Compounds into Connection Tables". *J. Chem. Doc.* **1967**, *7*, 165–169.

(14) Vander Stouw, G. G.; Elliott, P. S.; Isenberg, A. "Automated Conversion of Chemical Substance Names to Atom-Bond Connection Tables". *J. Chem. Doc.* **1974**, *14*, 185–193.

(15) A listing of some of these companies appeared in *Chem. Eng. News.* **1983**(Oct 17), 60.