LITERATURE CITED

(1) Balke, S., "Benutzerprobleme der Dokumentation und Information," *Nachr. Dok.*, **24**, 2 (1973).

(2) Cooper, W. S., "Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems," *Amer. Doc.*, **19**, 30 (1968).

(3) King, D. W., and Bryant, E. C., "The Evaluation of Information Services and Products," Information Resources Press, Washington, D. C., 1971.

(4) Meyer, R. L., Meskin, A. J., Mracek, J. J., Schwartz, J. H., and Wheelihan, E. C., "A Systematic Approach to Current Awareness and SDI," *J. Chem. Doc.*, **11**, 19 (1971).

(5) Skolnik, H., "The What and How of Computers for Chemical Information Systems," *J. Chem. Doc.*, **11**, 185 (1971).

(6) Stumpf, W., "Entwicklung und Erprobung von Methoden zur Auswertung in- und ausländischer Datenbänder für Retrievalzwecke im Bereich der anorganischen Chemie und ihrer Grenzgebiete," *Chem.-Zt.*, **96**, 301 (1972).

(7) Swets, J. A., "Information Retrieval Systems," *Science*, **141**, 245 (1963).

(8) Thompson, D. A., "Interface Design for an Interactive Information Retrieval System: A Literature Survey and a Research System Description," *J. Amer. Soc. Inform. Sci.*, **22**, 361 (1971).

(9) Wersig, G., "Zur Systematik der Benutzerforschung," *Nachr. Dok.*, **24**, 10 (1973).

Handling Commercial Product Names at Chemical Abstracts Service†

RUSSELL J. ROWLETT, JR.,* and DAVID W. WEISGERBER

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received January 11, 1974

Because Chemical Abstracts Service (CAS) abstracts and indexes a wide variety of technological and scientific literature, it is important that CAS be able to quickly and accurately equate the many commercial product names encountered in the literature with the actual complete chemical structures and the corresponding CA Index Names. This is accomplished readily within CAS processing by means of the CAS Chemical Registry System, a computer-based system that uniquely identifies chemical substances on the basis of their molecular structures. To the user of CAS products, the CA Index Guide provides a similar, although manual, link between the many commercial product names and the CA Index Names and Registry information.

Chemical Abstracts Service (CAS) publishes abstracts of the world's primary scientific literature which contains chemical and chemical engineering information and provides a variety of indexes to the original documents. All chemical substances for which new information is presented in the literature are indexed in *Chemical Abstracts* (CA) by name, molecular formula, and other indicators of structure. In order for a chemical substance name index to be useful, it must have all entries for a single substance appear reliably and consistently at one place in the index. Scattering of information in the index at synonymous substance names simply destroys the utility of the index because the user would never know whether he had located all references to a specific substance. This is particularly true of a large index such as CA, which now has over 630,000 individual Chemical Substance Index entries per six-month volume.

Equally important as having all index entries for a single substance appear at only one place in the index is the requirement that entries for related substances appear in proximity to facilitate generic searching. This ability to group related substances in an index is best accomplished by using fully systematic chemical substance names rather than their usual commercial or trivial names. Table I compares some common commercial and trivial names with the corresponding fully systematic names for the same substances. To obtain this reliability for its indexes, CAS has developed a comprehensive set of naming rules based on the nomenclature principles established by the International Union of Pure and Applied Chemistry.[1]

The problems arise when chemical substances are identified in the scientific and technical literature only by trivial or commercial product names such as Cinnamene or Dowanol EM, or perhaps only generic descriptions such as "the insecticide Gammexane" or "Polygard antioxidant." Placing such substances at their commercial product names in the CA indexes would simply scatter the chemical information. Such scattering of entries would make it almost impossible for a chemist searching CA to find all data for which he is looking. Just as a CA indexer must know where to place such substance index entries, so must the user of CA know where to look for such entries. Both require some way to rapidly and reliably equate the many commercial and trivial names from the original literature with the correct structure information and the CA Index Names.

Table II illustrates the type of problem an indexer, as well as a chemist preparing to search CA, might typically encounter. The five commercial product names are for the same common solvent. Some would be recognized immediately; others would probably not be recognized because they are less frequently used. But, for the purposes of indexing and searching CA, each of these names must be reliably converted to the CA Index Name "Ethanol, 2-methoxy-" and to the molecular formula $C_3H_8O_2$. The chemical substance name is inverted in the CA index; in normal text, this substance name will appear uninverted as "2-methoxyethanol."

While systematic nomenclature is essential to the production of an effective index, many chemists do not wish to become nomenclature experts. To facilitate their search of the CA indexes, the CA Index Guide identifies many commercial products and their CA Index Names. The CA Index Guide, which accompanies the CA Volume Indexes, is a collection of cross-references from the chemical substance

† Presented at the 166th National Meeting of the American Chemical Society, Symposium on Importance of Nomenclature of "Commercial" Chemicals in Chemical Safety, Chemical Disasters, and Chemical Literature, Chicago, Ill., Aug. 29, 1973.

* To whom correspondence should be addressed.

Table I. Trivial *vs.* Systematic Names

Benzocaine	Ethyl 4-aminobenzoate
Decalin	Decahydronaphthalene
Olamine	2-Aminoethanol
Sevin	1-Naphthalenyl methylcarbamate

Table II. CAS Recognition Problem

Commercial Product Names:

- Methyl Oxitol
- Dowanol EM
- Glycol monomethyl ether
- Methyl Cellosolve
- Poly-Solv EM

Reliably converted to:

- Ethanol, 2-methoxy-

Table III. Cross-References from Index Guide

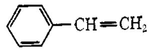
Lindane

See *Cyclohexane, 1,2,3,4,5,6-hexachloro-, (1α,2α,3β,4α,5α,6β)-* [*58-89-9*]

Acrylonitrile

See *2-Propenenitrile* [*107-13-1*]

Table IV. CAS Registry Record for Styrene

Registry number:	100-42-5
Structure:	
Molecular formula:	C_8H_8
CA Index Name:	Benzene, ethenyl-
Literature Names:	
Cinnamene	Styrene monomer
Cinnamol	Styrol
Phenethylene	Styrole
Phenylethene	Styrolene
Phenylethylene	Vinylbenzene
Styrene	Vinylbenzol

names used in the original literature, many of which are commercial product names, to the corresponding CA Index Names as well as cross-references from terminology used in other scientific disciplines and in commerce to the equivalent subject terms used in CA indexes. Notes which explain CA indexing policies are also included in the Index Guide. This data base has been derived from CA indexing practice over a period of many years. Table III illustrates some typical commercial product name cross-references appearing in the Index Guide.

For the purpose of CAS production, this rapid identification of a commercial product name is accomplished by means of the CAS Registry System, a computer-based vocabulary control tool used for the production of the CA indexes. The Registry System provides the important link between the chemical structure information, the CA Index Names, and the common commercial and trivial names appearing in the literature.

Because the CAS Registry System is basic to the operation of CAS and thus to the approach that CAS takes to handling the problem of reliably identifying commercial product names, it is necessary to review briefly the makeup and operation of this system.

The CAS Chemical Registry System is based on a concept suggested originally by G. Malcolm Dyson. It grew out of research in the early 1960's supported in part by the National Science Foundation. The CAS staff, building on work supplied by Du Pont, perfected an algorithm for generating a unique and unambiguous computer-language description of a chemical substance's two-dimensional structure and a means for adding stereochemical details to the computer record.[2] This algorithm became the foundation of a computer-based system that uniquely identifies chemical substances on the basis of their molecular structures and assigns to them an invariant, computer-checkable numerical identifier called a CAS Registry Number.

These Registry Numbers in themselves have no chemical significance. They are assigned in sequential order as the substances are entered into the CAS Registry System for the first time. Each Number designates only one substance and, therefore, furnishes an efficient means of substance identification independent of any system of nomenclature. This Number is used within the larger CAS processing system to link the structure with related files of substance names and other indexing data, abstracts, etc. Every specific chemical substance currently indexed in CA has an associated CAS Registry Number.

Pilot operation of the CAS Registry System began in 1965. Since that time, over 2.7 million unique substances have been recorded in the system. Substances new to the file continue to be registered at the rate of about 300,000 per year.

CAS Registry Numbers are in wide use today to provide exact identification of substances. They appear, of course, in CAS products; all specific CA Chemical Substance Index entries include the associated Registry Numbers. The American Chemical Society (ACS) primary publications, *Journal of Organic Chemistry* and *Inorganic Chemistry*, now include CAS Registry Numbers for those substances which are to be indexed by CA. The secondary publication, *Abstracts on Health Effects of Environmental Pollutants*, published by BioSciences Information Service of Biological Abstracts and comprised of selected material from BIOSIS and the National Library of Medicine's MEDLARS, includes Registry Numbers. More and more frequently Registry Numbers are being included as additional identification in chemical handbooks such as the Chemical Rubber Company *Atlas of Spectral Data and Physical Constants for Organic Compounds* and the United States Adopted Names (USAN) *Dictionary of Drug Names*. Information services, such as the National Library of Medicine's TOXLINE system, an on-line computer-based toxicology information service, are finding that CAS Registry Numbers provide a very useful search tool. These Numbers provide a common link among the several files that make up their combined data base; this is an invariant link that chemical nomenclature cannot supply because of the many trivial and systematic names, other than CA Index Names, that are in common use.

An important part of the CAS Registry System is the Registry Nomenclature File which supports the Registry Structure File. This computer nomenclature file contains the CA Index Name and all the commercial and trivial names which have been encountered in the original literature for each substance. There are approximately 3.6 million different names now on the nomenclature file. A large number of these are names other than CA Index Names and include the commercial product names.

Table IV illustrates a typical CAS Registry Record for a common commercial product, in this case, styrene. Styrene has been assigned the Registry Number 100-42-5; the last digit, 5, is a check digit which is assigned by a standardized computer calculation based upon the values and positions of the preceding digits. It is used by the computer to check the validity of the total Number each time the Number is entered into the processing system.[3]

A machine-language representation of the chemical structure for styrene forms the basis for the unique Registry Record. The molecular formula C_8H_8 also is recorded as a separate element of data and will form the basis for compilation of the CA Formula Index. The supporting Nomenclature File contains the CA Index Name and other commercial and trivial names for the registered substance. The

Table V. Index Entry Data Input

- Methyl Cellosolve
  detergents, for dry cleaning, 119428j
- Cythion
  skin penetration by, 128788d
- Freon C 318
  propellants, for shaving prepns., 67388z

Table VI. Index Entry Data Output

- Ethanol
  —2-methoxy- [*109-86-4*]
  detergents, for dry cleaning, 119428j ($C_3H_8O_2$)
- Butanedioic acid
  —[(dimethoxyphosphinothioyl)thio]-
  diethyl ester [*121-75-5*]
  skin penetration by, 128788d ($C_{10}H_{19}O_6PS_2$)
- Cyclobutane
  —octafluoro- [*115-25-3*]
  propellants, for shaving prepns., 67388z (C_4F_8)

systematic CA Index Name for Styrene is "Benzene, ethenyl-." In compiling the CA Chemical Substance Index and the Formula Index, the computer will select this name from the many on file and use it to format those index entries having the Registry Number 100-42-5.

Also shown in Table IV are the literature names that make up the rest of the Nomenclature File record for styrene. Any of these names, as will be shown later, can be used by an indexer for the purpose of retrieving the Registry Number 100-42-5 which, in turn, will provide the link for retrieving all of the important index data.

For normal indexing purposes, each Chemical Substance Index entry requires a Registry Number since this will provide during the final volume processing the key for retrieving all necessary index data from the Registry files. First, an attempt is made to obtain the Registry Number by name matching. If this fails, a structure match is attempted. If this also fails, the substance is new to the file and assigned a new Registry Number.

Accessing the files by structure requires that the entire structure be keyed into the system and then matched. The simplest and cheapest way to access the system is by name.

Thus, the CAS Registry System helps to support CA indexing by allowing the indexers to readily identify those chemical substances that CAS has previously indexed and for which index names have already been generated and structural information recorded. In this way, repetitive naming of the same substance is eliminated and those substances which are new to the Registry System and for which index names must be generated are identified.

CAS uses its Name Match system for the purpose of identifying commercial products by name in the following manner. When a chemist indexing a technical paper or patent encounters a commercial product for which an index entry will be made, he routinely inputs the substance name used by the author into the index data stream along with the indexing information. Table V illustrates how several typical commercial product index entries might be input. The indexer dictates his entries using the trivial names plus the text information and the appropriate CA reference identification. The material is then keyboarded into computer-readable form. The recorded names are subjected to a number of programmed edits and then searched by the computer against the Name Match File.

The Name Match File is a special file maintained by CAS for the initial matching of incoming names with names already on the Nomenclature File for the purpose of identifying Registry Numbers. The Name Match File is prepared by selecting names from the Registry Nomenclature File and condensing them by a special algorithm into a coded format. Incoming names such as those shown in Table V are condensed into the same coded format and then searched against the file. The reason for creating and searching against a separate file instead of using the Nomenclature File is that the Nomenclature File is a tape file sequenced by Registry Number. It is simpler and more rapid to search by name a file which is sequenced by name, albeit in coded form. The names are coded since this reduces the file storage to less than one-tenth of the space needed if the names were stored directly. Punctuation and capitalization are ignored in this file to eliminate name format variation and increase matching capability.

When these names are searched against the Name Match File and matches are found, the CAS Registry Numbers for these substances are identified and added to the index records. In turn, the CA Index Names and the molecular formulas are retrieved from the Nomenclature File and added to the index entries. Table VI shows the same index entries after the Name Match retrieval. The molecular formulas shown in parentheses do not appear in the Chemical Substance Index, but they are part of the index entry record and do form the basis for the corresponding Formula Index entries.

The CAS Name Match process now reliably identifies about 60% of all substance entries input for CA indexing. It has the important advantage of being quicker and less expensive than matching by complete structure input. Its use also eliminates the need for repeated identification by more than one indexer of the many commercial product names appearing in the literature. A name need be identified only once and then stored in the system for future reference.

When Name Match fails to identify the substance either because the particular name for an old substance has not been encountered previously in the literature or because the substance itself is entirely new, the structure record must be keyed into the system and matched with an existing registration or, if no match is found, added as a new registration. If the original document has not clearly defined the structure associated with a given name, then it also becomes a responsibility of the indexer to search the various reference sources in order to complete the registration. If the indexer can find no information about the chemical structure of a particular substance identified only by a commercial product name, the index entry will by default appear in the index at that name. If it can also be identified generically, then an additional index entry will be made at the class heading, such as Herbicides or Surfactants, in order to extend the index coverage.

All new names put through the Name Match process are added automatically to the Nomenclature File once Registry Numbers have been added to the index entries as a result of structure matching or new registration. In this way, the Nomenclature File is continually being augmented by new literature names.

There are certain computer checks and edits of the information present in and entering the Registry Nomenclature File. One such check that is important for maintaining the accuracy of the Name Match process is the identification of instances in which a substance name has been associated with more than one Registry Number. In a few cases, this situation is acceptable because some non-index names are so ambiguous that they can be associated with more than one registration (acronyms particularly fall into this category). In many cases though, this check serves to identify problems within the file. This situation does not prevent Name Match; it only gives multiple retrieval. But, this, in turn, requires resolution by the indexer and additional keyboarding to eliminate the incorrectly retrieved index information.

When a commercial product name fails to Name Match, but its chemical identity is known, it is added to the CA Index Guide data base as a cross-reference.

Because the Index Guide provides a manual link between the commercial product names and the corresponding CA Index Names and their CAS Registry Numbers, it serves as a somewhat limited printed version of the Name Match system. To ensure the accuracy of the Index Guide data base, the file has been checked against the Registry Nomenclature File. Because CAS uses a systematic chemical name whenever the chemical structure is known, the Index Guide cross-references may also be used to identify names or retrieve Registry Numbers.

In summary, the CAS Registry Name Match System provides a rapid and efficient way of equating by machine the many commercial product names encountered in a wide variety of technical literature with the systematic chemical substance index names and Registry Numbers to be used in the CAS products. In a similar way, the CA Index Guide provides to the user of CAS products a manual way to match many commercial product names with CA Index Names and Registry Numbers.

LITERATURE CITED

(1) CA Volume 76 Index Guide (1972), Section IV; Donaldson, N., Powell, W. H., Rowlett, Jr., R. J., White, R. W., and Yorka, K. V., "CHEMICAL ABSTRACTS Index Names for Chemical Substances in the Ninth Collective Period (1972-1976)," *J. Chem. Doc.*, **14**, 3-15 (1974).

(2) Morgan, H. L., "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service," *J. Chem. Doc.*, **5**, 107-113 (1965); Rowlett, Jr., R. J., and Tate, F. A., "A Computer-Based System for Handling Chemical Nomenclature and Structural Representations," *J. Chem. Doc.*, **12**, 125-128 (1972).

(3) CA Volume 76 Index Guide (1972), Section II, paragraph 10, p 71.

A Study of Citations to 308 Journal Articles in Chemistry Published in 1963

RICHARD W. GREFRATH

Mortvedt Library, Pacific Lutheran University, Tacoma, Washington 98447

Received March 26, 1974

A study was made of the behavior of citations to 308 journal articles in the field of chemistry published in 1963, and the citations to the articles were tabulated year by year to the present time (through the 1971 cumulation). The articles were classified as being practical or theoretical and by authors in college/university or in industry settings. The hypothesis was that theoretical articles would have a longer life than practical articles. Combinations of the various classifications of articles were made (*e.g.*, theoretical articles by industry affiliated authors *vs.* theoretical articles by college-affiliated authors) utilizing statistical analysis with the chi-square test. The results are discussed and evaluated.

PROCEDURE

To achieve the longest period of coverage, the journal articles were chosen from 1963 journal volumes, so that the citations to them could be looked up in all the yearly cumulations of the "Science Citation Index" (SCI) which began regular publication in 1964.

One of the shortcomings of the citation study by Brookes[1] is that consideration was not made of the increase in publication with each year. This defect is pointed out by Line,[2] who suggested that correction factors be applied to compensate for the increased obsolescence of the particular subject field, as well as the decay of the literature in the field. The desirable consideration of growth factor is stressed, which considers how many articles are published in successive years, but this is very difficult to determine for a particular subject field. A premise is that, if the likelihood of citation is the same for successive years, the earlier years will have fewer citations because there is less literature.

The problem is reflected in the SCI, in that the coverage of the SCI was increased each year; that is, the base of source journals from which citing articles were taken was expanded. In addition, the SCI revealed (in 1971) the results of the study in which the citation frequency varied from year to year, generally increasing by small amounts each year. Although this citation frequency is helpful, it does not consider in its computation the articles that were published that year but which were not cited at all; the figures merely reflect the citations per paper which appear in the SCI. These data are cited in Table I. This phenomenon presents a number of problems. It might seem an easy matter to simply adjust the citation data for each year based on the fraction of total coverage in that year. Although the calculations would be trivial, such manipulations would not really reflect the situation which the statistics suggest. In the present study, the field of chemistry was examined. Chemistry is an old discipline, and the journals for the field have been well established for a long time. Most of the additional coverage of the SCI reflects the birth of new journals which relate to newly evolved fields in science, for example, the whole body of new literature which has been generated by the popular interest in ecology. Further, new disciplines are created by increased specialization and merging of old disciplines, such as (some years ago) when biochemistry came into its own. As mentioned previously, the SCI coverage also extended to some areas of the social sciences, which have little effect on the field of chemistry. A list of the specific journals added was included in the 1971 cumulation: it contained no journals that seemed to be concerned with chemistry.[4] Also, many non-journal sources were added through the years, such as government reports and publications. So the mere yearly figures of "journals covered" is really not an accurate representation of the effect of the increased coverage in the field of chemistry, and it would be misleading and probably counterproductive to apply a corrective factor to the citations collected based on the coverage for that year. One possible procedure to account for the increase in articles would be to list all the journals which contain the *citing* articles in the 1964 cumulation, and then when counting citations in future cumulations only count those that appeared in the