# A Computer-Based System for Handling Chemical Nomenclature and Structural Representations*

RUSSELL J. ROWLETT, Jr.,** and FRED A. TATE
Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

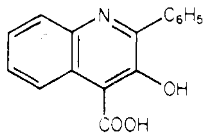
Received March 3, 1972

**The Chemical Registry System is reducing the professional effort required for indexing at Chemical Abstracts Service (CAS) and, at the same time, is providing a complete machine-searchable structural data base not previously available to the scientific community. CAS employs basic IUPAC nomenclature but conversion, beginning in 1972, to greater use of the fully systematic names, plus improved designs for the Registry files themselves, will lead to even greater usefulness of this system. Also, CAS intends to standardize the fundamental principles for naming cyclic structures so that procedures for the derivation of ring names can become more amenable to computer generation and translation.**

Ready and reliable access to scientific and technical information is becoming more and more dependent upon computer-based systems for building and searching large data banks and for organizing and composing search results in subject-oriented arrangements. This increasing dependence upon new processing technology eliminates many of the manual copying and editing steps which are inherent in traditional publication operations and considerably extends the usefulness and the useful life of the data which are managed through a well-designed information-handling system. The *Chemical Abstracts* (*CA*) Volume 71 Subject Index was produced by such a computer-based operation late in the spring of 1971. Earlier, each of the other corresponding *CA* Volume Indexes had been shifted to the same type of automated production. In the future, *CA* Collective Indexes also will be prepared by this system.

The automated system which Chemical Abstracts Service (CAS) is now using to produce all of its indexes utilizes computer-assisted keyboarding to reduce over-all input effort. It depends upon a single recording of each element of data, even though those data may later appear in several different contexts or formats in different indexes. The initial edit is accomplished by computer program so that only those items which do not pass this edit must be reviewed by staff. Also accomplished by computer program is the selection, from a closely coordinated set of nonredundant, computer-oriented processing files, of the contents for each type of index. The indexes are then organized, formated, and composed automatically. Because the data are thus transferred along the processing chain without the traditional intervention of human transcription, most of the need for galley-proof is eliminated. The computer-based production system is, therefore, quicker, cheaper, and more reliable than the previous traditional publication process. As indicated above, an additional bonus is that, since the data are also recorded in machine-readable form, the entire content of the publications is searchable by computer program. Such searching can provide responses to questions which are not answered easily by human search of corresponding large printed indexes. And, as we will discuss later in this paper, the data base can be analyzed to yield statistics that are unavailable in the manual system and which aid in the refinement of the system.

The specific topic of this paper is the handling of *CA* index names for chemical substances within the over-all CAS index production system, including the CAS Chemical Registry System which controls this indexing vocabulary. The point of the present discussion is the difference between the nomenclature system of the International Union of Pure and Applied Chemistry (IUPAC) and *CA* index nomenclature which has, in general, been derived from the IUPAC system. IUPAC nomenclature provides a systematically derived textual equivalent of the corresponding structural diagram; each IUPAC name is a "stand-alone," unambiguous representation of the analogous molecular structure. However, despite the need for a single, invariant name of a substance for use in large, systematic indexes, such as those of *CA*, the IUPAC system usually provides two or more equivalent name alternatives. For instance, Figure 1 shows three alternative IUPAC names for the given structure. Each of these alternative IUPAC-approved names emphasizes a particular portion of the total structure. The preferred name is the second one since it focuses on the carboxylic acid function



4-carboxy-3-hydroxy-2-phenylquinoline
3-hydroxy-2-phenyl-4-quinolinecarboxylic acid
4-carboxy-2-phenyl-3-quinolinol

Figure 1. Systematic names

*As presented to the XXIII Conference of the International Union of Pure and Applied Chemistry, Boston, Mass., July 1971.

**To whom inquiries should be addressed.

which has priority, according to an established hierarchy, over the other chemical functions represented.

The IUPAC hierarchy of chemical functions is not, however, all inclusive. Particularly when there is a choice of more than one accepted nomenclature system, as there is for such compounds as heterocyclic amines, azo compounds, and ethers, the IUPAC practice has been to accept all systems rather than to establish preference for one. In contrast, a highly systematic vocabulary with no unrecognized synonymy in the terminology is necessary to produce a reliable, useful index with predictable, consistent positioning of entries. In a discipline-wide chemical index, the largest single class of vocabulary terms consists of the names of substances.

The *CA* index name meets all of the requirements of the corresponding IUPAC name but, in addition, is defined so as to bring the names of structurally related substances into juxtaposition in an alphabetical index listing. In cases in which the IUPAC system provides more than one possible name, CAS follows a more comprehensive hierarchy so that the same index name will be assigned consistently. For this purpose, CAS has greatly extended IUPAC rules and developed other special ordering rules. Last revised in 1967, this nomenclature system comprises three to four times as many rules as the IUPAC system on which it is based. The additional explanatory memos required for special chemical subjects have numbered 459 since the *CA* index nomenclature practices were documented in 1967.

Volume 71 Subject Index statistics show that the handling of substances for just a single six-month volume of *CA* requires approximately 183,000 different index terms. Figure 2 illustrates the composition and derivation of the index name of the same quinoline compound used in the previous example. That portion of the *CA* index name which precedes the comma is called the "heading parent." There are 31,000 "heading parents" in Volume 71. The index name of a substance used once in CAS operations is permanently maintained in the vocabulary in case the corresponding substance appears again, although actually only a small portion of the substances reappear in the next *CA* volume or within the next year. Vocabulary control must be stringent. Without the use of a single, invariant name for each substance and without the ordering of similar substances within a large index collection, it would be almost impossible to locate data from the nearly 1.3 million primary documents which will be covered in the *CA Eighth Collective Index* for the 1967-71 period.

The computer-readable version of the *CA* Subject Index is providing statistics which have never before been available and which continue to illustrate the problems of a controlled vocabulary for substances. Of the 31,000 heading parents in Volume 71, for example, 46% are listed only once. That these single-entry headings represent substances new to the literature is a logical assumption, but this assumption is not borne out by the CAS Chemical Registry System. Since the Registry System became operational at the beginning of 1965, it has recorded nearly two million different substances—each represented by a different *CA* index name—but only 30% of the substances now entered are new to the Registry files.

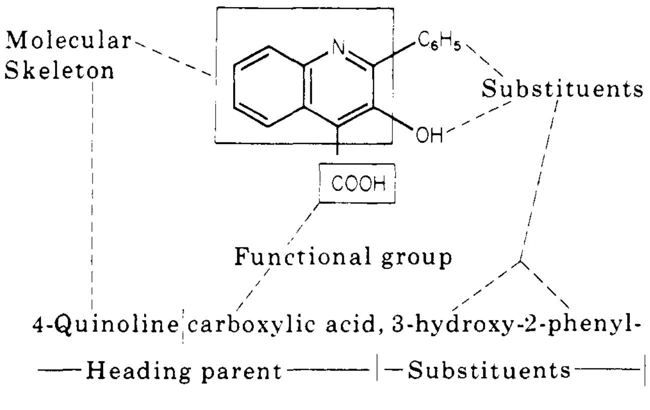


Figure 2. Vocabulary used in systematic names

As noted previously, only a little over 50% of the total known substances have been included in the published studies of two or more groups of investigators. Fewer yet have become a part of applied science, readily recognized in the primary scientific and trade literature by name alone. All other substances are defined in terms of source, elemental composition, or qualitative structural descriptions—i.e., structural diagrams. Moreover, most substance names which do appear in the primary literature either are not derived systematically or represent simple structures for which systematic names are derived easily. Thus, most systematic names are derived solely for the purpose of listing them in indexes such as the *CA* Volume and Collective Subject Indexes.

CAS does not advocate that systematic nomenclature be used without exception throughout the primary literature. Such a practice would be very expensive, would slow the primary publication process, and would yield a large number of incorrect names. Training in nomenclature use, which requires a minimum of twelve to eighteen months even for a highly experienced chemist, is not practical for all chemist-authors and editors. So, for most common substances, unsystematic and trade names will continue to appear in the primary literature. Such descriptive names as diolamine and tetralin, for example, are useful in the trade and primary literature, but these common substances appear with the ethanols and naphthalenes, respectively, when included in a highly organized general index such as the *CA* Subject Index. The commonly used unsystematic names also are listed in the index. Only by cross-referring these to unambiguous, consistent names, derived by the same set of rules used to generate the names of similar substances, however, can the index become so highly structured that all information about similar derivatives of the same heading parent is collected at the same general location in the index.

Another reason for including trivial or unsystematic names in the *CA* Indexes is that these indexes do not serve only chemists. No other branch of science provides fully systematic textual equivalents of the structural diagrams employed throughout science and technology as the basic qualitative identification of substances. It follows that many substance names—most of them not based on the structure of the substance—which have attained a special value within a given science must be cross-referred in *CA* subject indexes.

In the traditional manual system of building a chemical index, each substance to be indexed required the drawing of the corresponding structural diagram. This procedure was repeated each time a primary paper, patent, or report included significant data on that substance. With the exception of a few very common substances—less than 2% of those identified in a given *CA* volume subject index—the drawing of the structural diagram was necessary for the generation of an index name. Each name was then recorded on a separate index card along with the corresponding bibliographic citation; when the indexing was complete, the cards were alphabetized. Under the manual system, just the alphabetizing of the more than ten million index entries expected in the *CA Eighth Collective Index* would have required a minimum of eighty man-years. In the alphabetization step, the duplicate entries were identified and merged, but at this point in the procedure, the intellectual effort of naming all of the duplicate compounds had already been expended. In addition, since CAS makes multiple use of some of the index data in Volume and Collective Subject, Formula, and Ring Indexes, the same substance name is used a minimum of six times. Thus, in the traditional system, many manual checks of the character string and reference were required each time an entry was copied for one of these uses.

The CAS Chemical Registry System circumvents the regeneration of many index names by automatically comparing unsystematic or trade names that appear in the primary documents against a large, carefully edited file of such names. If a match is obtained, the corresponding *CA* index name and molecular formula are retrieved automatically from the computer file. If no match is obtained, or if the primary document includes no name for the substance selected for indexing, the corresponding structural diagram is entered directly into the computer and likewise matched against the 1.8 million structures already on record.

The name match procedure is being implemented so that no index name will need to be generated more than once. Also, the recording of an edited and verified, ready-for-use name within the processing files eliminates the need for double checking each time the name is retrieved from the files for use in a different index. But these two advantages do not assure economical and timely indexes for the future. As indicated previously, almost 200,000 different substance names are included in a current six-month *CA* Volume Subject Index, and very few of these names recur in the next successive six-month index. Also, the Registry records show that at the present time about 300,000 new substances are added to the files each year. As the new structures required to be drawn become more and more complex, the derivation of the corresponding index names becomes increasingly more time-consuming and expensive. The task is made even more difficult by the widespread use of trivial names.

IUPAC nomenclature has traditionally used many so-called trivial chemical names as heading parents and to these heading parents has added systematically-named substituents. The result is a semisystematic name which results in much confusion when employed in a highly structured index such as *CA*. Relatively simple derivatives of trivially-named hydrocinnamic acid, for example, are cross-referred successively through a series of acceptable trivial names, leading in the end back to hydrocinnamic acid. Beginning with the ninth collective index period in January 1972, CAS will assign to this and similar structures, names that are more systematic.

Figure 3 is illustrative of some of the changes CAS will be making in the shift from nonsystematic IUPAC names to fully systematic index names. In the ninth collective index period (1972–76), the nonsystematic names for all but a handful of very familiar compounds, such as acetic and benzoic acids and phenol, will be carried only as cross-references in the Index Guide. In addition to the simplification of indexer training, this change will provide two important benefits. First, it is a step essential to developing computer programs which will automatically generate *CA* index names from structural diagrams entered into the computer store via typewriter or other input device and directly recorded in computer-readable form. Second, the fully systematic names will improve the searchability of the computer-readable version of the Volume and Collective *CA* Subject Indexes by providing easier recognition of the systematically named functional groups and ring systems.

| As named in the CA 8th Collective Index | To be named in the CA 9th Collective Index |
|---|---|
| Acetone | 2-Propanone |
| Anisole | Benzene, methoxy- |
| Crotonophenone | 2-Buten-1-one,1-phenyl- |
| Glyceric acid | Propanoic acid, 2,3-dihydroxy- |
| Oxamide | Ethanediamide |
| Vinyl alcohol | Ethenol |

Figure 3. *CA* index names

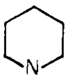
By 1972 CAS will be automatically checking each new index name added to the Registry Files against the corresponding structural diagram which is added separately to the computer files. This program is already operational. It is planned that when funds become available, substances identified in pre-1965 *CA* subject indexes will be entered into the Registry Files by these same programs without the need to prepare corresponding structural diagrams and to record them separately. Such programs will thus conserve professional staff effort. It is only through such computerized aids that CAS can reasonably plan to continue its broad support of the information needs of chemists into the 1980's when, estimates indicate, more than one-half million abstracts per year will have to be prepared and indexed.

The CAS Registry System accepts structural diagrams recorded in computer-readable form and automatically converts each into a unique matrix, checking against the file to determine whether that substance has been registered previously. The actual computer record is very significantly compacted and coded so that the structure records for a quarter of a million substances are contained on one 2400-foot reel of magnetic tape (9-track 1600 bpi). The reverse process, the generation of an acceptable structural diagram from the unique record stored in the computer file, is more difficult. But this capability is being developed, and such an output system will be operational within the next few years. There is, of course, a direct correlation between the automatic generation of systematic index names and the automatic generation of structural diagrams. Specific identification of substructures, including functional groups, and skeletal features, such as rings, is necessary whether the fragments are to be named or displayed in diagrammatic form.

Of the 1.8 million unique structures currently entered in the Registry System, 86% contain ring moieties; included are almost 27,000 different ring systems. Presently, the Registry System is being adjusted to identify individual ring structures, specifically, within a total molecular structure. By the end of 1972, each ring system included within a molecule will be identified automatically as each new structure is registered. If a ring system is listed in *The Ring Index*, the reference will be supplied automatically, thus aiding the indexer in the generation of the total *CA* index name. A novel ring will be directed automatically to an expert in ring nomenclature for the assignment of ring locants and a ring index name, and for the recording of the new structure in *The Ring Index*. These capabilities will provide the basis for automatically generating an Index of Ring Systems as a part of regular index production. Automatic generation of the ring locants and ring index names is a longer range goal; it now appears that such development will require considerable simplification and standardization of the present international systems for naming ring structures.

Problems of ring names permeate almost the whole of chemical nomenclature and are compounded by the proliferation of ring-naming systems over the years. There are at present at least seven approved IUPAC systems for naming rings, and four more either are partially accepted or are being considered for acceptance. The use of these systems yields many approved names for the same ring or for combinations of one ring with other rings. An example is seen in Figure 4 which shows the variations in naming the pyridine ring as it occurs in combination with other rings. It is presently impossible to search *CA* Subject Indexes for all forms of such rings. They can be found by careful, time-consuming search of the Index of Ring Systems.

Nomenclature options for ring systems
containing the following:



Piperidine
Pyridine
Indolizine
Pyrindine (2 isomers)
Isoquinoline
Naphthyridine (6 isomers)
Quinoline
Quinolizine
Quinuclidine

Acridine
Phenanthridine
Quindoline
Quinindoline
Thebenidine
Acrindoline
Numerous "indicated hydrogen" forms
Various von Baeyer structures, e.g.,
1-Azabicyclo[4.1.0]heptane

Figure 4. Ring nomenclature

| Number of Rings | Number of Basic Ring Systems | Percentage of Sample |
|---|---|---|
| 1 | 1,661 | 6.60 |
| 2 | 3,653 | 14.53 |
| 3 | 6,070 | 24.15 |
| 4 | 4,836 | 19.24 |
| 5 | 4,000 | 15.91 |
| 6 | 2,230 | 8.87 |
| 7 | 1,180 | 4.69 |
| more than 7 | 1,502 | 6.01 |

Figure 5. Ring system frequency analysis

There are now ring frequency analyses which show the number of rings in the total collection of parent rings contained in CAS files; as shown in Figure 5, three-membered rings comprise the largest percentage of the known basic ring structures. Of the individual rings, the six-carbon ring has the highest frequency of occurrence. The pyridine ring, used as the example in Figure 4, is the heterocyclic ring that occurs most frequently.

Thus, the Chemical Registry System serves as a vocabulary control system which not only deals with chemical names and the corresponding structural diagrams interchangeably but also assures that a given substance is always represented in the *CA* Volume Subject Indexes by the same index name; as pointed out earlier, statistics such as those above on the content of the CAS indexes are available only because of the computer data base.

# A Multilingual Index Via the Multiterm System*

HERMAN SKOLNIK
Hercules Incorporated, Research Center, Wilmington, Del. 19899

Received February 16, 1972

**The Multiterm system, which was introduced by the author in 1970 as a new indexing method, is shown to be a unique concept for producing a multilingual index via computer processing. An inherent advantage of the multilingual Multiterm index is that on reaching a certain size, the translating of terms from a source language to others becomes essentially a computer operation. The new concept has significance for international cooperative programs.**

There has been considerable interest in the potential for worldwide information systems, particularly for various disciplines of science and areas of technology. Most recently, this interest was expressed in some detail in the Unesco publication "UNISIST," with many cogent arguments.[1]

Despite the apparent feasibility documented for a viable international information system, it is highly unlikely that real progress will be made towards its realization until solutions are found to two basic problems:

1. The linguistic barrier.
2. A universally accepted system for indexing.

The objective of the work described in this paper was to find solutions to these two problems. Input and output character limitation of our computer restricted the work to languages based on the Latin alphabet.

## THE MULTITERM INDEX

The Multiterm index is a system for communicating the informational content of documents by coordination of subject terms in defined directional orders.[2] It was conceived

*Presented before Middle Atlantic Regional Meeting, ACS, Philadelphia, Pa., Feb. 16, 1972. Hercules Research Center Contribution No. 1584.