Pure & Appl. Chem. Vol. 49, pp.1781 -1792. Pergamon Press, 1977. Printed in Great Britain.

THE CAS DATA BASE

Ralph E. O'Dette

Chemical Abstracts Service, Columbus, Ohio, U.S.A.

Abstract - All information collected by Chemical Abstracts Service (CAS) is input to a single data base from which all CAS printed publications and computer-readable services are derived. As a guide to users of these publications and services, this paper briefly describes the content, structure, and management of the CAS data base. The sources, selection, and processing of documents; the use of authority data files such as the CAS Chemical Registry System in data base building; the output of selected data elements in various media and formats; and the value of CAS-produced user aids are discussed.

The title of the symposium implies that retrieving information may be useful, important, and perhaps even interesting; however, my assignment is not to deal with retrieving or using information but to clarify the preparation of one important information collection from which useful information can be retrieved.

It is as though I had been asked to open a Cordon Bleu cooking class by lecturing on cattle-grazing or growing artichokes. One can become a rather good cook, if not a great one, in spite of abysmal ignorance about the production of one's raw materials. One can also find information in spite of abysmal ignorance about the modern techniques and systems that are being used today to help make the information retrievable, but to attempt to search today with yesterday's underlying assumptions about the search process and about the files being searched betrays a dogged determination to be ineffective.

The information user does not have to become expert in the electronics of computers nor even in the linguistics of their programming in order to use modern information transfer technology. But the user should have a general comprehension of the outlines of modern information processes in order to exploit them. To use another analogy, highly effective use of quite sophisticated analytical instrumentation is made by people who could not build nor repair, let alone design the instruments.

My assignment is to describe how we capture information at Chemical Abstracts Service (CAS) for our publications and services.

The information user will find it profitable to look at CAS differently today than five or ten years ago. CAS is one of a number of information services that have been quite dynamic over the past dozen years, and one result has been to place much greater search power in the hands of their users.

The CAS data base is the product of intellect, supported by computer technology. Documents are selected for analysis by subject specialists; abstracts and index entries are prepared by subject specialists. Progress is being made in providing those specialists with electronic and mechanical aids, but the intellectual content of the CAS data base is intellectually derived.

Some jargon will ease the balance of my discussion, specifically, the phrases "data base" and "data element". "Data base" has a precise meaning among some specialists, but common usage has it as simply a more-or-less coherent collection of information. There is some preference for the term "information base" in order to preserve the word "data" to mean numerical information, but "data base" seems firmly entrenched.

A "data element" is a precisely and consistently defined unit of information. It has a specific content, set of characters, and arrangement. A given data element may be whatever the system designer and editor choose - - an author's last name or an entire citation, a chemical element or a molecular formula - - but the value of the concept is in the consistent use of one definition for a given data element, once chosen. CAS data elements have been designed to encompass the smallest units of information that users may want to call for or that we may need to manipulate in building a publication or other information service.

Application of automated editing in building the data base is also greatly facilitated by the ability to deal with quite limited, discrete pieces of information.

All of the information that is found in any CAS publication or service passed through the CAS data base during production of that publication or service. Every bit of information is either a data element or part of one.

A discussion of the CAS data base can be approached from several points of view. The ensuing discussion starts at the data base, and works towards input and output. Figure 1 is a very generalized flow diagram, to place the data base in context. During the input phase, primary literature is received, selected, and analyzed and the analysis is input to the computer. These processes create the computer-readable data base. Output processes select data elements from the data base to produce a variety of publications and services. The familiar Chemical Abstracts with its semi-annual volume and collective indexes represents virtually the entire data base. Other CAS services created in the past 15 years represent an emphasis on some subset of the total data base.

THE CAS SYSTEM

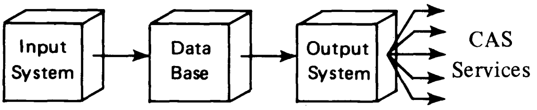


FIG 1.

TARGET SYSTEM

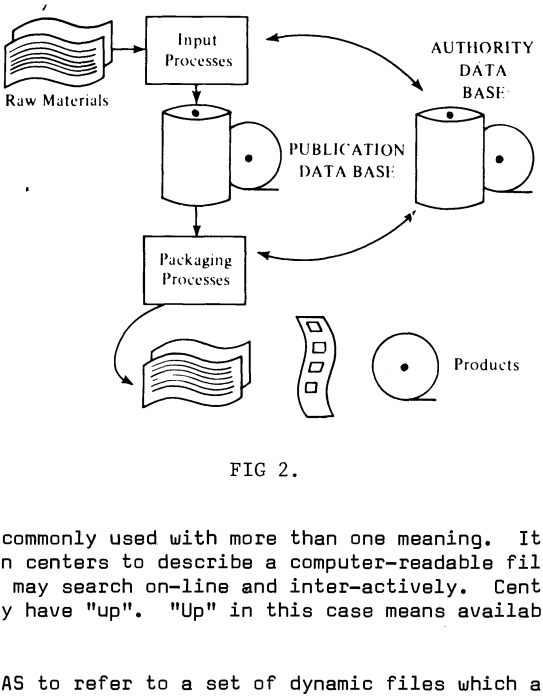


FIG 2.

The term "data base" is commonly used with more than one meaning. It is a cliche that is current among information centers to describe a computer-readable file which they will search for users or which users may search on-line and inter-actively. Centers often speak of the various "data bases" they have "up". "Up" in this case means available for searching on the system.

"Data base" is used at CAS to refer to a set of dynamic files which are the next-to-last step in the manufacture of services. The set of data base files consists of two sub-sets -- the publications data base and the authority data base (Fig. 2). The publications data base contains the bibliographic identifications, abstracts and index information that we derive from source documents; the authority data base provides vocabulary control and other types of processing support. Both data bases are also sources of search aids for users. It is essential to stress that "authority" in this sense means our self-imposed authority over what we permit ourselves to do. There is no intent to imply that these files are authoritative over the outside world.

It is useful to visualize the CAS publications data base as having three strata (Fig. 3)

According to this image, one layer contains all of the information needed to identify the primary or source documents that come into the system. That descriptive information is referred to as "bibliographic data". Another layer contains the texts of abstracts. A third layer contains the information from which the subject indexes will be compiled. Various extracts from one or more of the layers constitute existing CAS services.

The bibliographic data answer such questions about the document as: Who wrote it? Where and when was it published? Where was the reported work done? What language was it published in? There are some 60 such pieces of information that might be useful simply to identify a particular document in its role as a carrier of information, and these are what go into the bibliographic data component of the data base. There are almost three dozen more document-related data elements such as the abstract number and the CA Section Number. All of these "data elements" are uniquely identified and consistently defined, and each may be separately addressed and retrieved as may be desirable to become part of an information service.

The abstract text component of the CAS data base is, basically, a description of what the authors of the source documents did and why they did it. Years ago, when the world was a more leisurely place, some CA abstracts were, effectively, brief papers. A chemist could take such an abstract into the lab as a guide to repeating a synthesis, for example. Today, because of the great and continuing growth of the literature and because of a refinement of our view of our proper role in the world of information, we try to make the abstract serve in the same capacity as the lead paragraph on a well-written newspaper story. It contains the essence of what, why, when, where, and how. That much information will satisfy many searchers to whom the article is of passing interest, but it will not seriously impede those to whom it is of no interest. At the same time, we also try to make it enough to enable other searchers to decide whether the full article is worth pursuing.

DATA BASE INFORMATION CONTENT

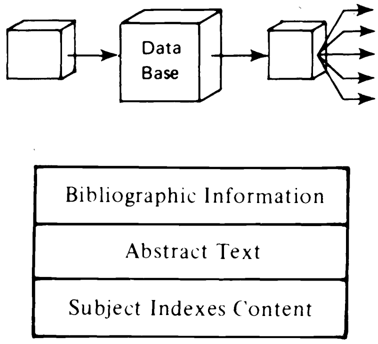


FIG 3.

CAS SUBJECT MATTER INDEX DATA ELEMENTS

Concept Heading (CTH)
Functional Category (CAT)
Qualifier (QLF)
Heading Parent (PAR)
Homograph Definition (HOM)
Line Formula (LIN)
Substituent (SUB)
Name Modification (MOD)
Stereochemistry (STE)
Molecular Formula (FOR)
Registry Number (REG)
Text Modification (TMD)
Primary Publication Type (PUB)
Abstract Number (CAN)

FIG 4.

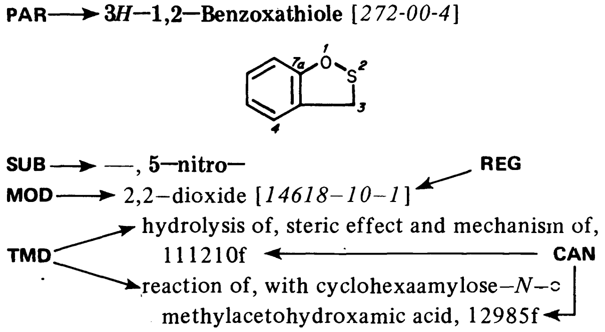
The subject indexes component of the data base is the most complex, in its structure and in its use as a source of services. There is no need to convert CAS users into authorities on how the CAS processing system operates, but some knowledge of how the data base is assembled will prove quite valuable to the user who wants to get the most he can from CAS services. Whereas with the bibliographic data it is apparent that the names of authors and journals, publication dates, languages, and so on, help to specify a document and thus guide the user to find it, the data element structure of a subject index entry provides an additional level of interesting and useful insight into its intellectual structure.

Up to 14 different data elements may be used by an analyst to comprise entries for the CA subject matter indexes. Figure 4 lists the 14 possibilities with their three-letter codes. Figure 5 shows a CA Chemical Substance Index entry and identifies the six data elements from which that entry was built.

A CA Formula Index entry is shown in Figure 6 with its data elements identified.

Users of CAS publications and services frequently are interested in names of substances. To define completely the CA index name of a chemical substance requires up to five of the possible six data elements. Figure 7 highlights those of the subject matter index data elements that pertain specifically to index names of chemical substances, a topic that will be dealt with again below.

To summarize to this point, the CAS data base is the product of intellectual analysis supported by computer-based technology. The data base is, in a sense, a large tank into which are steadily poured those small units of information, the data elements, which describe a primary document and its chemically significant content. It is a flow-through container, because as new information pours in, information services are drawn out. The analogy is also misleading, however, because taking data elements out does not physically remove them. In the aggregate, however, the CAS data base tank represents the current literature of chemistry and chemical engineering. Because the data elements are small, 'the nature of what is to be withdrawn from the tank in the form of user service may be quite precisely defined.


PAR → *3H*—1,2—Benzoxathiole [*272-00-4*]

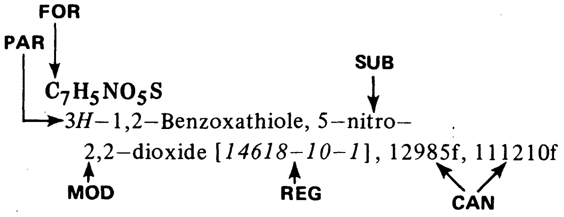
SUB → —, 5—nitro—

MOD → 2,2—dioxide [*14618-10-1*] ← REG

TMD → hydrolysis of, steric effect and mechanism of, 111210f ← CAN

TMD → reaction of, with cyclohexaamylose—*N*—methylacetohydroxamic acid, 12985f ← CAN

FIG 5.


FOR → $C_7H_5NO_5S$

PAR → *3H*—1,2—Benzoxathiole, 5—nitro— ← SUB

2,2—dioxide [*14618-10-1*], 12985f, 111210f

MOD, REG, CAN

FIG 6.

CAS SUBJECT MATTER INDEX DATA ELEMENTS

Concept Heading (CTH)
Functional Category (CAT)
Qualifier (QLF)
Heading Parent (PAR)
Homograph Definition (HOM)
Line Formula (LIN)
Substituent (SUB)

Name Modification (MOD)
Stereochemistry (STE)
Molecular Formula (FOR)
Registry Number (REG)
Text Modification (TMD)
Primary Publication Type (PUB)
Abstract Number (CAN)

FIG 7.

ABSTRACTS BY
TYPE OF SOURCE
(1975)

69%	Periodicals
17%	Patents
8%	Conference papers
2%	Technical reports
2%	Dissertations
2%	Books
100%	

FIG 8.

The user, it would seem, is most interested in what he can get out of any given information system. One of the major limits on what he can get out is what is put in by the system's operators.

The first step in building the CAS data base is to collect source documents. A document is a source document if it fits stated criteria of form and content. Figure 8 lists the types of source documents abstracted by CAS and the percentage of CA abstracts that each contributed in 1975.

Figure 9 shows the largest national sources of abstracted papers in 1975, based on the addresses of first authors; for comparison, the percentage of journal paper abstracts by language of publication of the paper is also shown. In 1975, abstracted journal papers represented R & D performed in 137 different countries, and the papers were originally published in 51 different languages.

In 1975, 454,000 documents were identified to the data base. 392,000 were abstracted. An additional 62,000 proved to be equivalent patents, and so they were cross-referenced to the first patent received on that invention. For 1976, about 450,000 documents will be cited.

It is also pertinent to note the growth pattern of the data base. Figure 10 shows numbers of documents cited since 1907, the first year of publication of CA. Figure 11 shows such data for the past 25 years; while growth is not smooth, the trend of the curve is steady.

With respect to content, CAS policy is to cover chemistry and chemical engineering comprehensively. To make that general guideline explicit, we publish a 182-page (plus index) definition of subject coverage, entitled "Subject Coverage and arrangement of Abstracts by Sections in Chemical Abstracts." (Fig. 12). The 1975 edition is current. The coverage manual serves a number of useful purposes for users, one of which is to make specific the CAS operational definition of chemistry and chemical engineering. While the manual is specific to CA, CA and the CAS data base are virtually identical with respect to subject scope as well as to data element content. The exception to that statement is our business-oriented service, CHEMICAL INDUSTRY NOTES, which is not dealt with in the coverage manual.

WORLDWIDE NON - PATENT LITERATURE
(1975)

National Origin		Pub. Language	
U.S.	25.8%	English	59.7%
U.S.S.R.	24.6%	Russian	23.3%
U.K.	6.3%	German	4.8%
Japan	7.3%	French	3.0%
Germany	6.8%	Japanese	3.0%
France	4.1%	All other	8.2%
All other	25.1%		

FIG 9.

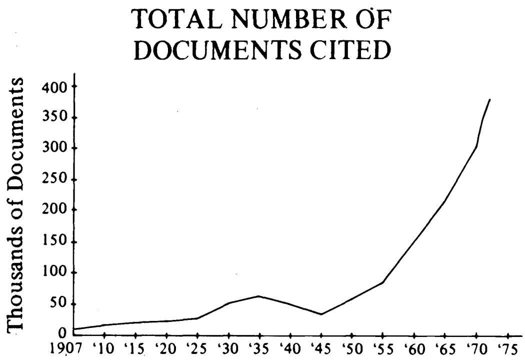
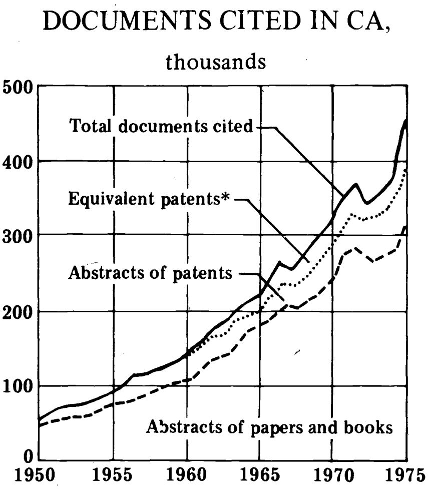
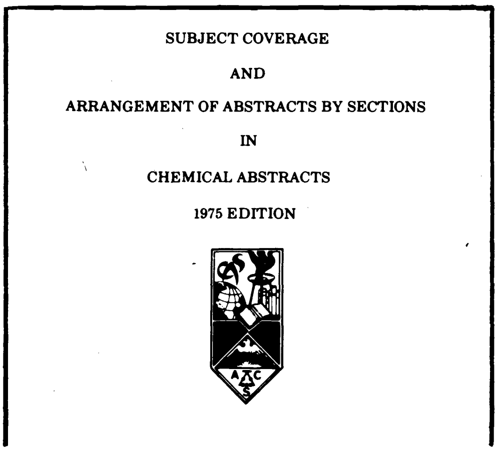


FIG 10.



* Covers inventions already patented in another country. These are cited in *CA Patent Concordance*.

FIG 11.



SUBJECT COVERAGE

AND

ARRANGEMENT OF ABSTRACTS BY SECTIONS

IN

CHEMICAL ABSTRACTS

1975 EDITION

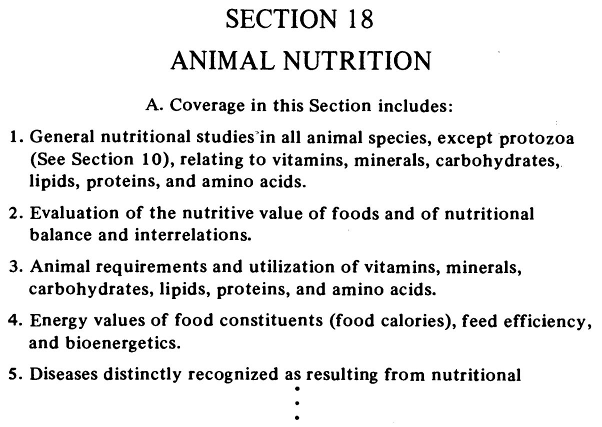
FIG 12.

Figure 13 is an excerpt from a typical page from the coverage manual, in this case showing part of the detailed definition of CA Section 18, Animal Nutrition. Figure 14 shows another part of that definition; the asterisks call attention to several non-chemical aspects of animal nutrition which are recognized as closely related to CAS coverage but which are not covered. Others of the 80 sections of CA are defined in lesser or greater detail, depending on the need to define a boundary between chemistry and a related but different discipline. For example, the scope of CA Section 27, "Heterocyclic Compounds with One Hetero Atom", can be defined in many fewer words than can the obviously multidisciplinary field of animal nutrition for which CAS coverage cannot be self-evident but must be specified in terms of what is excluded as well as what is included.

In CAS internal jargon, processing begins with source document packages collected by our library. An issue of a journal or the complete volume published from a conference is a document package. Technically trained staff select from such packages the individual papers - - these are the actual source documents - - that will be analyzed for the data base.

These first steps in assembling the CAS data base have been stressed because one of the foremost criteria of the usefulness of an information service must be its content. There is room for doubt whether many information service users have thought very deeply about the published raw materials from which the data bases they use are assembled. Since we are quite systematic about selecting documents for our data base, some knowledge of our ground rules should prove useful to users.

Figure 15 summarizes the selection process statistically. In 1975, the CAS library passed on to the selection group 3,200,000 potential source documents. The selectors kept only 1 of 7 as appropriate to the CAS data base. The total selected is the total number of documents cited, a statistic noted above.



SECTION 18

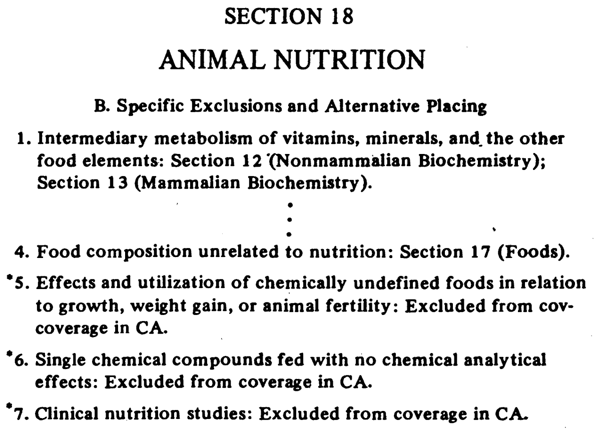
ANIMAL NUTRITION

A. Coverage in this Section includes:

1. General nutritional studies in all animal species, except protozoa (See Section 10), relating to vitamins, minerals, carbohydrates, lipids, proteins, and amino acids.
2. Evaluation of the nutritive value of foods and of nutritional balance and interrelations.
3. Animal requirements and utilization of vitamins, minerals, carbohydrates, lipids, proteins, and amino acids.
4. Energy values of food constituents (food calories), feed efficiency, and bioenergetics.
5. Diseases distinctly recognized as resulting from nutritional

⋮

FIG 13.



SECTION 18

ANIMAL NUTRITION

B. Specific Exclusions and Alternative Placing

1. Intermediary metabolism of vitamins, minerals, and the other food elements: Section 12 (Nonmammalian Biochemistry); Section 13 (Mammalian Biochemistry).

⋮

4. Food composition unrelated to nutrition: Section 17 (Foods).

***5. Effects and utilization of chemically undefined foods in relation to growth, weight gain, or animal fertility: Excluded from cov-coverage in CA.**

***6. Single chemical compounds fed with no chemical analytical effects: Excluded from coverage in CA.**

***7. Clinical nutrition studies: Excluded from coverage in CA.**

FIG 14.

SOURCE DOCUMENTS
(1975)

3,200,000	Screened
454,245	Selected
392,234	Abstracted as new
62,011	Cross-referenced as equivalent patents

FIG 15.

Once the individual source documents are chosen, each is identified to the computer system i computer-readable form. As noted above, the inventory of bibliographically useful data elements that may be recorded for all types of source documents approaches 100. On the orde of 10-15 different items of information are captured to identify the typical source document Examples are the article title; authors; journal title, volume and number; location where the reported work was performed, and publication language. Many of the other data elements are required specifically for the CAS Source Index, used for locating copies of source documents not in a user's library as well as for interpreting the unusual citation.

These items of information identify the source document without saying much, if anything, ab its content, unless, of course, the title is highly informative. At this point, while information is being added to the publications data base, the authority file component of th data base has also begun to function. The input keyboard operators do not enter titles of journals, for example; they enter a five-character code - - a CODEN* - - plus a computer-calculated check character to detect transcription errors. At output time, an authority file will be called on to translate the CODEN into a standard journal title abbreviation.

When the subset of bibliographic data elements selected for each source document is recorded each data element with its identifying code is associated with a temporary abstract number. The temporary number serves to corelate the source document with the data elements derived from it throughout its processing until a permanent number is assigned for publication in Chemical Abstracts or another service.

Next, the source document is sent either to a staff analyst or to one of the part-time abstractors around the world who contribute towards building the CAS data base. The assignment is made by subject specialists who know the subject and language skills of each analyst

The analyst, who has language skills as well as subject competence, can derive up to 40 data elements in preparing his abstracts and index entries. He dictates his analysis of the source document into a tape recorder, prefacing each statement or bit of information with the appropriate data element identifier. The story is, in fact, much more complex: some of the analyst's input consists of various signals which trigger the automated system to supply still other data elements, to add to authority files, and so on. But the number 40 is not a misleading indicator of the fineness of the analysis.

The further support that the authority data base supplies at this point may be illustrated b two rather different examples, one concerning index headings for the CA General Subject Inde and one indexing nomenclature for chemical substances.

Figure 16 lists the kinds of entries that are found in the CA General Subject Index. Figure 17 illustrates the structure of General Subject Index entries, which consist of main heading under which are alphabetized words or phrases that specify the aspect of the main heading dealt with in a specific abstracted source document. There are at present some 47,000 main headings resident in a computer-readable authority file and available for use in this index; 36,000 of these headings are taxonomic terms dealing with biological entities. The file is called on during automated editing of index data elements. In its present application, the analyst will be informed via computer printout if he inadvertently uses an unapproved heading. Under some circumstances, the editing program will make corrections, such as from the singular form to the plural or vice versa. Figure 17 also shows that when there are man entries under a heading, the heading is sub-divided. A recent volume index to CA had two columns of entries under each of the first two of the headings, a column under each of the next two, and so on.

The aim of the index heading authority file is consistency in the application of index headings. Work is going forward to apply the authority file-type of control to the modifyin text phrases also, but this application is being approached very cautiously. Consistency is virtuous but not at the risk of seriously reducing the informativeness of the index entri

* American Society for Testing and Materials (ASTM) Standard E250-76

CONTENT OF CA GENERAL SUBJECT INDEX

- ● Processes
- ● Reactions
- ● Properties
- ● Physical phenomena
- ● Apparatus
- ● Plant and animal headings
- ● Plural and class headings [compounds (alcohols), minerals, fossils]
- ● Naturally occurring materials (air, coal, sand)
- ● Commercial commodities (spandex fibers, gasoline)

FIG 16.

Illustration GSI Heading Subdivisions PHYSIOLOGICAL ORGANS AND TISSUES

Eye
acetopropyl alc. effect on, 143348m
Eye, composition
cornea
sorbitol dehydrogenase of, 51067r
Eye, disease or disorder
penicillin pharmacol. in, R 100322f
Eye, metabolism
of arsenic, in rainbow trout, 1021a
Eye, neoplasm
retina, rhodopsins of, 76485p
Eye, toxic chemical and physical damage
from chloroquine, 61758f

FIG 17.

which are intended to achieve maximum specificity. Work is also going forward on giving the analyst on-line access to such authority files from a terminal so that he may do his edit checking immediately.

And, while information retrieval is not the subject of this paper, it may be in order to point out that users may obtain a complete listing of the General Subject Index headings. Such a list is available as part of a package of search aids for the computer-readable CA Subject Index Alert.

The second authority file example to be described is the CAS Registry System. It is useful to stress again the narrowness of this use of the word "authority". The word is used strictly in the information processing sense. An authority file does not tell users what they may or may not do; it tells CAS staff what we may or may not do based on prior decisions about such matters as definitions of terms and various editorial practices.

The primary function of Registry is the exercise of vocabulary control over the indexing of substances, but it has many corollary functions, as well.

One objective of CAS in indexing substances is to provide a reliable link among the various kinds of nomenclature used in the chemical and related worlds. While reflecting the various ways that authors may name substances, including various systematic, non-systematic, and trade names, it is important to prevent uncontrolled scatter of entries about substances in the indexes themselves. To accomplish these ends, what is required is a comprehensive collection of names from the literature plus an infallible technique for recognizing to the limit of possible precision what is intended by a name. The CAS Chemical Registry provides all these capabilities.

To illustrate, suppose that a substance to be indexed is a completely defined labelled isomer of a particular steroid that is identified in a journal paper by a structural formula and a descriptive but non-chemical name. Registry techniques can deal with all of the defined molecular detail in the structure. Registry will, for example, distinguish that isomer from an unlabelled isomer with the same stereochemistry or from another labelled isomer with one different stereo center. A CAS nomenclature specialist will generate a CA indexing name if the right one is not on file, and the name that the author used will be filed as a synonym.

In using Registry, the CAS analyst identifies in each source document those names used by the author of the document for all substances that CAS indexing policy encompasses. These so-called "author names" - - substance names used by authors - - are then compared by the computer system with a file of all names that CAS has encountered since the system began in 1965; there are presently well over 5 million names on that file. The computer may report that the candidate name is reliably synonymous with a CA indexing name. In that case, a molecular formula and Registry Number are supplied by the computer, and the Registry Number stands in for the index name throughout further processing. When the abstract is printed, the author's nomenclature is used in it; when the index entries are printed, the reference to the abstract will be under the CA index name. Cross-references between authors' names for substances and CA index names appear in the Index Guide, about which more below.

About half of the time, the process described above occurs: the substance name used by the author of a paper is recognized by the nomenclature phase of the CAS Registry System. If the name match process does not produce a match, a chemist must draw a structural formula based on information from the source document. The structural formula is input to the computer on a chemical typewriter, and it is automatically compared with a computer file of structural formulas, now numbering about 3.6 million. About 10-15% of the time the name will not match but the structure will; this indicates a new name for a known substance. About 25% of the time, the structure is new to the file, and so it automatically receives a new Registry Number, and a CAS nomenclature specialist derives the CA indexing name. This information is then added to the authority files and is available for computer match the next time the names or the structure appear in the literature. There have been about 350,000 such new substances per year in recent years. Most of this Registry activity is the automatic response of the CAS system to the use by a staff analyst of simple signals as he identifies substances to be indexed or to be highlighted in abstracts, or perhaps both.

In summary, then, the CAS Registry System exercises vocabulary control by means of a system consisting of (1) highly systematic indexing nomenclature; (2) techniques for mechanical input of all of the molecular structural detail reported by an author, including stereochemistry, labelling, unusual valence, and so on; (3) an algorithm that assures that duplicate structures cannot accidentally reach the file because uniqueness is always recognized; and (4) computer authority files of structures and names that encompass every substance that CAS processes.

Since at CAS it is a matter of principle that any part of the information processed may be considered a resource for service to the public, the authority files take on a new character. Since these files represent, in effect, implementations of editorial policy, when they can be made available to CAS user audiences in whole or in part they can be powerful aids to more effective searching. For example, a massive cross-reference from CAS Registry Numbers to CA index names has been published, and another such cross-reference, from so-called common names to Registry Numbers is being experimented with. The full Registry Structure file as a computer-searchable file is also being actively studied by a number of groups in Europe and the U.S. The Registry files are singularly powerful examples of the information processing concept of authority data bases.

So far, this discussion has aimed to project the idea of the CAS data base, what it is, what it contains, and how, by examples, it is constructed and managed. The next few paragraphs concern the third box in Figure 1, and illustrate, by example again, how the data base contents are converted into information services, including the generation of various aids to the use of the output services.

It has been stated several times that all of the information in the CAS data base and, therefore, in any CAS publication or computer-readable service, exists in the form of specific, defined data elements. The data base, then, is a parts inventory for any service to be produced. In practice, a service is defined as the sum of its parts. If the service is to be printed, the format and type size of each element of the service is specified, and the data element identifiers become composition codes for the strings of symbols they encompass. Computer programs then utilize the data element identifiers to direct a computer-driven photocomposer, and the composer delivers fully-composed pages, ready for the printer's processes. None of the information necessary to fix the typography or layout of any element of a publication actually resides on the data base. That information is supplied by the program that guides the final packaging of a publication or information service.

This approach to composition also makes it possible for a given data element to appear in many different contexts and in widely different typographies appropriate to those contexts. The technique is one that helps mightily to meet the objective that information people have in mind when they speak of "single input-multiple outputs". Figures 5 and 6 illustrate that point.

The data base conceot in the preparation of secondary or accessing information services makes possible a subsidiary but valuable extra benefit. To treat the topic of user aids adequately would require a full-length paper. What follows is necessarily a summary (Fig. 18).

CAS USER AIDS

- Subject Coverage Manual
- Introductions in publications
- Index Guide
- SDF Specifications Manual
- Special search aids
- User education program

FIG 18.

The Subject Coverage Manual has been discussed. Its value in understanding the subject scope of the data base cannot be overemphasized.

For those particularly interested in computer-readable services and particularly for the computer specialist who must manipulate these services, a three-volume loose-leaf SDF Specifications Manual defines the standard distribution format, defines each CAS computer-readable service in terms of its data element content, and defines each data element.

Each printed publication from CAS includes in the first number of each volume a detailed introduction to the contents and structure of that publication. For example, the author index to each CA volume notes a number of the peculiar problems inherent in mixing names of organizations and authors from dozens of countries and languages. CAS encounters 500,000 or more names of authors each year and must rationalize the many different national customs and language characteristics. The CA author index explains 24 procedures that are followed.

The computer and the data base also make possible a potentially infinite array of what are called "special search aids". The value of these aids in the selection of terms for search questions, whether the search is by computer or is manual, is probably self-evident. Fig. 19 illustrates a word-frequency list. If a searcher uses commonly occurring words, he increases the likelihood of many and of irrelevant answers. If the searcher insists on rarely used words in the interest of getting a few exact answers, he may miss pertinent responses. The word frequency list is also a useful guide to the occasional misspellings and typographical errors on the file. Figure 20 illustrates a Key-letter-in-context or KLIC Index. Note how the KLIC index calls attention to similarities among words of different meanings.

CA CONDENSATES SEARCH AID
WORD FREQUENCY LIST

318	THERAPEUTIC
1	THERAPEUTICA
12	THERAPEUTICAL
12	THERAPEUTICALLY
1	THERAPEUTICALS
48	THERAPEUTICS
1	THERAPIC
1	THERAPIES
1319	THERAPY

FIG 19.

CA CONDENSATES SEARCH AID KEY LETTER IN CONTEXT

	THERAPEUTICALLY	12
	THERAPEUTICS	48
CHEMO	THERAPEUTICS	17
	THERAPY	1319
PHARMACO	THERAPY	29
RADIO	THERAPY	175
CHEMO	THERAPY	485
IMMUNO	THERAPY	40
E	THERATE	79

FIG 20.

CA INDEX GUIDE

- Synonyms
  - Substance names
  - Other terms
- "see also" Cross references
- Indexing policy notes
- Introduction to Subject Indexes
- CA Nomenclature System

FIG 21.

The final user aid on the list is the CA Index Guide. (Fig. 21). This unique and powerful tool would be unthinkable without the computer-readable data base and the steady flow through it of the vocabulary of current science and technology. The Guide includes an alphabetical list of synonyms for substance names as well as other terms. It contains index cross-references and indexing policy notes. The Guide also contains a section on the organization of CA subject indexes, and, finally, it contains an extensive summary of CAS policies and practices for section of indexing names for chemical substances. At the calculated risk of oversimplifying, the Index Guide is a translator between the free, varied, and constantly changing language of the scientific and technical public and the calculated and conservative precision of CAS index language. All attempts to exploit the CAS data base must begin in the Index Guide unless there is a premium for carelessness.

The purpose of this discussion has been to describe the structure of the CAS data base as a guide to the more effective use of publications and information services produced from it.

Further information on some aspects of this topic will be found in the publications (1-8).

REFERENCES

1. D. B. Baker, Chemical & Engineering News 54, 23-27 (1976).
2. D. L. Dayton and A. Zamora, Journal of Chemical Information and Computer Sciences (in press).
3. P. G. Dittmar, R. E. Stobaugh, and C. E. Watson, Journal of Chemical Information and Computer Sciences 16, 111-121 (1976).
4. C. R. Gustafson, J. D. Rule, G. G. Vander Stouw, and C. E. Watson, Journal of Chemical Information and Computer Sciences (in press).
5. L. J. O'Korn, Proceedings of the Symposium on Algorithms for Chemical Computation (in press).
6. Chemical Abstracts Service, Chemical & Engineering News 53, 30-38 (1975).
7. Chemical Abstracts Service, CAS Report 4 (1975).
8. Chemical Abstracts Service, CAS Report 5 (1976).