Improved methods for the preparation of an index from marked printers' proof have been developed. These methods are adaptable to the conventional process using 3 × 5 cards or the more sophisticated data processing techniques.

A generalized computer program for index preparation could probably be justified by publishers as a service to authors. This would assure rapid production of uniform index copy and save personal and secretarial time for both authors and editors.

ACKNOWLEDGMENT

The authors wish to express thanks to Professor L. F. Fieser for permission to publish his improved method of index assembly in this paper. Thanks are also due to K. H. Zabriskie of The Chemical Abstracts Service for his help in initiating the preparation of our index, to E. I. du Pont de Nemours & Co. for their cooperation in the project, and to Gessner G. Hawley of the Reinhold Publishing Co. for making it possible to use a new type of index in the third edition of "Formaldehyde". We are also indebted to Reinhold for permission to reproduce portions of the index in this paper.

LITERATURE CITED

(1) *Chem. Eng. News*, **42**, 62 (Oct. 19, 1964).
(2) Fieser, L. F., private communication (Sept. 29, 1964).
(3) Fieser, L. F., Fieser, M., "Organic Chemistry," 3rd ed, Reinhold Publishing Corp., New York, N. Y., 1956.
(4) "Organic Syntheses," Collective Volumes, John Wiley and Sons, Inc., New York, N. Y.
(5) Walker, J. F., *J. Chem. Doc.*, **4**, 45 (1964).
(6) Walker, J. F., "Formaldehyde," 2nd ed, Reinhold Publishing Corp., New York, N. Y., 1953.
(7) Walker, J. F., "Formaldehyde," 3rd ed, Reinhold Publishing Corp., New York, N. Y., 1964.

Processing, Publishing, Storing, Correlating, and Retrieving Biochemical Information at Chemical Abstracts Service*

KENNETH H. ZABRISKIE, JR., and MICHAEL F. LYNCH
Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received July 7, 1965

Recent developments at CAS in collecting, publishing, storing, manipulating, and retrieving biochemical information in a computer-based system are reviewed. The relation of this system to other operations at CAS is shown.

I. INTRODUCTION

This paper describes how biochemical information is handled at Chemical Abstracts. In particular, it describes a newly developed operation, the "Chemical-Biological Activities Information System" (called CBAC), and compares it to the way CAS deals with biochemical information in the regular issues and indexes of *Chemical Abstracts* (CA). An explanation of the methods used in CBAC for chemical compound identification and for vocabulary control is followed by a discussion of how CBAC will function both as an alerting service and as a retrospective search program.

II. OBJECTIVES OF THE CBAC INFORMATION SYSTEM

The primary purpose of establishing the CBAC System is to provide rapidly concise and well-indexed information of chemical-biological interest, exclusive of the plant kingdom.** Equal to the primary purpose is the aim of accumulating from the same analysis steps a store of organized chemical-biological activity information, suitable for retrospective searching from a wide variety of viewpoints.

* Presented before the Divisions of Chemical Literature and Medicinal Chemistry, Symposium on Drug Information, 148th National Meeting of the American Chemical Society, Chicago, Ill., Sept. 1, 1965.

** Note Added in Proof. The plant kingdom has been included in CBAC coverage since Jan. 1, 1966.

The CBAC System will in time include both a biweekly publication and a computer search facility. The publication began in January 1965; the search service should be ready early in 1966.

III. COMPARISON OF CA AND CBAC

Let us examine and compare some significant features of CA and CBAC. This comparison of the scope, manner, and format in which CA and CBAC deal with biochemical information will prove useful, both as a review and as a framework for more detailed discussions to follow.

1. Number of Abstracts (Digests) per Year
   - CA: 55,000 abstracts (1964) in 19 sections, ranging from "Enzymes" to "Fertilizers" from "Foods" to "Radiation Biochemistry."
   - CBAC: Approximately 10,000 digests in 1965, in the single area defined as biological activity of organic chemical compounds and the effects of biological systems on organic chemical compounds.
2. Abstractors
   - CA: Voluntary workers in the biochemical field, keeping themselves and their co-workers abreast of the field.
   - CBAC: Permanent staff of full-time analysts, trained by CAS and working with close guidance.
3. Controls on the Form of the Abstract (Digest)
   - CA: Kept within general guidelines set up by CAS for abstractors followed by careful editing.
   - CBAC: More consistent format made possible by the fact that the area being covered is well defined, fairly specific, and written by CAS staff.
4. Indexing
   - CA: Trained permanent staff of indexers and index editors, operating separately from, and later than the abstractors and abstract editors.
   - CBAC: The intellectual analysis step generates the index simultaneously with the preparation of digests.
5. Vocabulary Control

   Chemical Compounds
   - CA: Author's nomenclature is used in the abstracts; CA nomenclature is used in the indexes; no machinable record is produced.
   - CBAC: Authors' nomenclature is used, supplemented by CA nomenclature where needed for clarity; other nonproprietary names may be included. The names and structures are stored in machinable form.

   Other Vocabulary
   - CA: Abstractors have free choice to use the authors' words or to substitute their own; indexers use a controlled vocabulary.
   - CBAC: The same vocabulary is used in both digest and index; open-ended and carefully edited.
6. Availability of Indexes
   - CA: Keyword Index in each issue–a terse phrase; the full Subject Index for a volume appears about six months after the volume ends. Five-year collective indexes begin to appear after the close of the collective period and the publication of the final volume indexes for the period.
   - CBAC: Keyword-in-Context Index in each issue; six-month cumulative indexes require two weeks to prepare and are published after the period ends. Cumulative indexes may be run off at any time for any period desired. Specialized indexes for selected topics can be prepared at will within a few days.
7. Number of Index Entries per Abstract (or Digest)
   - CA: About 12 to 15 access points per abstract, including Subject Index entries, cross references, and Author and Formula Indexes.
   - CBAC: About 25 to 30 on the same basis.

   - CA: Reading abstracts in sections of interest in each issue; manual search of Keyword Index in each issue; manual search of Volume and Collective Indexes.
   - CBAC: Rapid scanning and/or manual search of each issue; machine search of accumulated information; scanning of machine-produced lists, prepared on demand in areas of special interest.

IV. CONTROLS ON VOCABULARY

In the comparisons above, there is frequent indication of controls in the CBAC System which are either not present or not as strict in the regular CA operation. Let us examine the nature of the control methods which are important to the successful operation of CBAC.

As a consequence of selecting a particular area of biochemical information, we were able to establish several broad controls on the language. These controls, in turn, enable the analysts to produce sentences which are readable, meaningful, and capable of being stored in standardized machine language. At the same time, the material is suitable for programmed rearrangement, searching, or listing, in a myriad of ways, as present or future users' needs may dictate. It is in this area of flexible and rapid response to changing patterns of user needs that the CBAC Information System best illustrates the power of mechanized information handling.

Some of the controls used in the CBAC System are as follows:

A. Control of compounds by Registry Number (described below), name, or structural formula.

B. Vocabulary controls on

1. "Modification" and "Function": *via* an open-ended, but carefully edited list of terms describing responses and reactions.
2. "Target": The names selected for the organs and systems within the host are in accordance with "AMA Standard Nomenclature of Diseases and Conditions."
3. "Host": Species names and family names follow accepted classifications of the animal kingdom.

C. Occasionally the information in the article can only be expressed accurately in the sentence structure by using additional qualifying words. Such modifying phrases are permissible; the vocabulary of these additional descriptors is open-ended, but carefully edited.

As the CBAC store of information accumulates, and the vocabularies grow, it is likely that the organized list of terms in use will be published, perhaps with data on frequency of occurrence.

V. REGISTRY SYSTEM–REGISTRATION–REGISTRY NUMBER

Full understanding of the topic of this paper requires at least a condensed and simplified description of the Registry and of Registry Numbers.

The CAS Research and Development Division has developed, with the assistance of many groups of people within and outside of CAS, a system for assigning a unique

Registry Number to each compound of known structure. The registration process consists of entering the atom and bond connections into the computer, converting to a standard form by machine, and testing the result against previously registered compounds stored in similar format. As a result, the machine either (a) locates the Registry Number (if the compound has been entered before) or (b) assigns a new Registry Number (if the compound has not yet been registered).

The fact that the Registry System computer programs store all the atom and bond information in a standard connection table makes it possible to search the file either (a) for specific compounds (as in registration) or (b) for substructure units. The latter search, often called a generic search, is a powerful new chemical information tool which arises from mechanization of information systems. Finding all members of a set of compounds which have the same substructure units present and/or absent is literally impossible today unless one undertakes a page-by-page search of all CA indexes.

VI. USE OF CBAC AS A CURRENT AWARENESS AND ALERTING TOOL

Let us turn now to the current awareness and alerting function of the system. It should be emphasized again that there is but a single intellectual analysis step involved in processing biochemical information into the CBAC store. The knowledge is stored in a manipulatable form, which allows us to develop the resultant awareness information tools and services in the frequency, formats, and fashion best suited to the users' needs.

The CBAC system was initiated with a new biweekly publication, CBAC, inaugurated with the issue of January 11, 1965. The style, format, coverage, frequency, and indexing of this new journal were determined finally only after prolonged user study. Discussions and meetings began in early 1962, with major review meetings at the Atlantic City ACS Meeting in the fall of 1962, the New York ACS Meeting in September 1963, and the Chicago ACS Meeting in September 1964. Four different samples were produced to obtain user review and comment for our guidance.

The CBAC publication consists of:

1. Digest Section
2. Keyword-in-Context Index
3. Molecular Formula Index
4. Author Index

The Digest Section is arranged by journal, and the title of the journal appears in abbreviated (but directly readable) form at the head of each list. Digests are consecutively numbered and contain:

1. The names and Registry Numbers (where applicable) of the compounds in the article.
2. Sentences which describe the biological activity (or chemical activity of biological interest) reported in the article. Sentences are consecutively numbered within digests.
3. Structural formula drawings for many registered compounds.
4. Journal CODEN, volume, and page references to the original article.

The names of compounds begin at the left edge of the printed page and the Compound Registry Number appears in the left margin next to the name. Important conceptual terms are printed in capital letters.

The Index Section is composed by:

1. Permuting the sentences in the body of each CBAC Digest and in the title of each CBAC article.
2. Combining the permuted sentences and titles in one list.
3. Alphabetizing by keyword.

In the Index Section, each alphabetizing word is in capital letters, as is each word in the title and each capitalized word in the Digest. Each index line refers to a digest number or sentence "sub-number."

The Author Index cites digest numbers for each author represented in an issue, and the Molecular Formula Index cites the digest and the Registry Number of the compound in each issue.

The entire publication is "type-set" on a modified 1401–1403 computer-printer, using a set of 120 characters, including both upper and lower case letters and superscript and subscript numbers.

VII. RETROSPECTIVE SEARCHING IN THE CBAC SYSTEM

The intellectual organization performed on the information before it is processed into the CBAC computer record makes it possible to manipulate the knowledge in the store in a wide variety of ways. (However, the searching techniques described below will not become operational until the CBAC System inventory grows to useful size in early 1966.)

Some of the possible search routes are as follows:

1. If the search is structure-based we may enter the system *via* one or more of the following paths: (a) name of a compound (or compounds), (b) structure of a compound (or compounds), (c) partial structure, or a set of substructure units (including "but not" logic).

2. If the primary search variables are conceptual, we may enter the system by one or more of these routes: (a) host animal, (b) host genera, (c) portion of the host studied, (d) body function affected, (e) nature of the change in function, (f) other factors present in test animal, (g) route of administration.

3. A correlative search is one that combines two or more of the foregoing parameters.

As an example of a structure-based search using substructure units, consider the following. A biochemist has need to test a new compound of known structure for several possible areas of activity. He and his team have postulated (a) that certain portions of the molecule may cause the desired activity, and (b) that another portion of the molecule will cause difficulty (such as poor solubility or unwanted side effects). It is desired to search the CBAC records for biological activity information on compounds which contain the former moieties and not the latter. The Registry System and substructure programs provide the Registry Number of all compounds in CBAC which meet these criteria, and the CBAC System will provide the digest reference numbers in which any biological activity of these compounds is discussed, or the sentences themselves, if desired. If the list of candidate

compounds is long, or if the questioner wants to focus his attention on a particular type of activity, or host, etc., additional search parameters may be selected from the conceptual area. Combining search parameters in this fashion is an example of "correlative searching" based on structure.

As another example of a search (one which is primarily *via* the verbal or conceptual information in the store), consider a research man who is interested in learning what drugs or compounds have shown anticonvulsant activity when administered to small mammals infected with tularemia. This search can produce as an answer any or all of the following: (1) pertinent digest reference numbers, (2) Registry Numbers and names of compounds with the activity specified, (3) the text of the sentences in the digests which meet the question criteria.

VIII. ROLES OF THE ANALYST AND OF THE MACHINE IN THE CBAC SYSTEM

The main functions of the scientific literature analysts are:

1. To examine an article and select or reject it for CBAC (following a course cut by the Assignment Department staff on a "fail safe" basis).
2. To extract from accepted articles the pertinent information concerning biochemical activity within the CBAC frame of reference.
3. To construct one or more meaningful sentences describing the information from (2).
4. To re-examine the article briefly, and to determine whether the digest accurately reflects the content.

The analysts use standard input forms to record their digests. The digest are then keypunched, including the codes which indicate the sentence parts.

The computer programs in the CBAC System perform the following steps:

1. Organize the information in the individual digests into the format desired.
2. Construct the permuted entries in the Keyword-in-Context Index section and organize in alphabetical order, with appropriate use of capital letters.
3. Tabulate the Author Index.
4. Locate existing Registry Numbers for compounds already in the Registry.
5. Assign Registry Numbers to compounds not previously registered.
6. Verify the given molecular formula for each compound by calculation from the atom-bond table and tabulate the Molecular Formula Index.
7. Printout a master copy of the four sections of the CBAC issue as input for the photography and printing operations.

Associated steps in the CBAC work flow plan include: structure drawing and molecular formula determination (by chemists working from names); clerical listing of atom-bond relationships in the drawn structure; and key-punching the atom-bond lists. The clerical operations are also internal parts of the Registry operation described in section V of this paper.

IX. SUMMARY

In addition to its direct usefulness to the biochemistry field, CBAC represents an important step forward at CAS for several other reasons:

1. CBAC is the first operational system developed by CAS R&D which will be capable of providing direct answers, rather than references, in response to queries.

2. CBAC begins to close the gap between chemical structure information systems technology (which is well along) and conceptual information processing techniques (which are still in the developmental or even research stage).

3. CBAC combines digest and index production in a single intellectual step and will be a valuable pilot study for assessing the long-range usefulness of such a combination to CA issues and indexes.

4. CBAC is a significant move toward the over-all objective of the CAS R&D program, namely: smooth integration of modern, computer-based information processing methods with the traditional CA issues and indexes, modified as deemed necessary and appropriate.

The result will be a unified chemical information center with a variety of services and publications, coordinated with other centers and with major users, *via* a network of communicating channels carrying input, questions, answers, and, we hope, the satisfied comments of our users.