Computer-Based Composition at Chemical Abstracts Service*

WILLIAM C. DAVENPORT and JOHN T. DICKMAN
Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received August 10, 1966

The Chemical Abstracts Service is in the process of converting its abstract and index handling procedures from a strictly manual operation with hot and cold typesetting processes to a fully integrated computer-based process. All abstract issue and index services will be available both in traditional printed form and in a machine-processable store useful for alerting, retrospective searching, and reorganization. The basic conversion will be complete by 1970.

Chemical Abstracts (CA) began publication in 1907. The main purpose of this information service has been to make the results of chemical and chemical engineering research, development, and applied technology accessible, usable, and used. Until 1961 this objective was met solely through the conventional publication of volumes of printed abstracts and indexes. Since 1961 Chemical Abstracts Service (CAS) has broadened the means of providing information to include specialized computer-based services for current awareness as well as for treating selected subject areas in depth. In addition to production of computer-based services, a store of information in a form suitable for computer manipulation is also being accumulated. Through use of the computer, this information can be adapted to the varying needs of individuals and organizations, and in-depth retrospective searching is made possible. CAS activities are currently being broadened to include additional computer-based alerting and retrieval services, and the potential for customized retrieval services is being greatly strengthened.

This paper describes, in part, the approach taken by CAS to integrate on a computer base all of its operations, from information acquisition through publication and provision of retrieval services. Notable progress has been made, such as by MEDLARS (*1*), in utilizing computers in an integrated system to produce a periodical, special bibliographies, and mechanized retrieval services. The CAS system under development utilizes many principles in common with MEDLARS but differs in the requirement for a greater number of re-uses of input data.

Because of the degree of integration that exists in the new CAS system, the title of this paper is not completely descriptive. "Composition" in the context of publishing generally refers to the several steps of typesetting, proofing, make-up, etc., between a manuscript and final printing press runs. In this conventional sense—the placing of printed characters on a page—composition is only one phase (the output) of the CAS system. Composition in the present sense at CAS is much broader. It begins with the receipt of the serial publications, patents, reports, and books which are CAS input. It is at this early point that treatment of the information within the computer-oriented system includes steps which are in a significant sense inseparable from the remainder of the processing system which culminates, in part, in printing.

The CAS system to be described in the following pages centers around storage of standardized, hardware-independent, representations of characters as well as of information "modules" such as titles, authors, chemical compounds, and the like. Translation programs have been written so that input in a wide variety of character sets from a wide variety of devices will be translated internally to the standard sets which can then be output, again through a translation technique, in a wide variety of forms—*e.g.*, magnetic tape, punch card, film, printed paper—by a wide variety of devices. Thus, for example, a "module" of information, such as the title of a paper, need be input only one time although it may reappear as output in various forms in a variety of CAS publications and services, some as yet undesigned. The alpha-numeric symbols that constitute the "module" are also completely described in storage so that a full range of printing flexibility may be used when appropriate equipment is available but so that, also, translations will accommodate the more limited equipment available today.

The CAS system is thus compatible with but is also independent of present-day conventions and equipment and is "open ended" so as to take advantage of more advanced forms of equipment which inevitably will arise.

Furthermore, since all CAS data base files are recorded in this standard form, complete compatibility is achieved among all data base files. The problem of data base obsolescence as a result of new advances in input, computing, and printing hardware is therefore avoided.

That the volume of chemical literature has been growing at a rapid rate is well known. Table I shows how *CA* has grown to keep pace. In 1950, 57,600 abstracts appeared in *CA*. It is estimated that in 1966, approximately 215,000 abstracts will be published. That growth represents an

* Revised version of paper presented before the Division of Chemical Literature, 151st National Meeting of the American Chemical Society, Pittsburgh, Pa., March 23, 1966.

Table I. Growth of *Chemical Abstracts*

Year	Volume (No.)	Abstracts	Entries	
			Subject Index	Formula Index
1950	44	57,600	281,500	71,000
1955	49	85,600	521,000	149,000
1960	54	132,000	716,500	199,000
1965	62–63	195,000	1,170,000	350,000
1966	64–65	215,000[a]	1,300,000[a]	400,000[a]

[a] Estimated.

annual increase of 9% per year, compounded. A corresponding increase in index entries will result in approximately 1.3 million Subject Index entries and 400,000 Formula Index entries for 1966.

The number of characters required to compose CAS publications is also large. Composition of the *CA* Formula Index requires 300 different characters, the Author Index 400, the Subject Index 500, and *CA* issues require 1450 different characters (*2*).

The massive volume of information and the need to maintain discipline-oriented coverage dictate the need for specialized services and computer composition. In order to continue effectively the conversion of CAS from a strictly abstracting-indexing-publication operation to a comprehensive computer-based information service, it will be necessary to integrate those procedures that require intellectual analysis and to mechanize handling the analyses and composing the text so as to provide a full range of services in the most economical manner.

To date, *CA* has been composed by conventional processes which, characteristic of all traditional printing processes, present the reader with two serious problems. First, conventional publication processes, of indexes in particular, restrict the publisher to one hierarchy and a single hierarchy leads inevitably to buried information; and,second, conventional publication processes restrict the re-use of information because each re-use requires repeating much of the same effort which went into the first publication. If, as CAS intends, a comprehensive, versatile, and adaptable chemical information system is to be built and operated at reasonable cost, it is necessary to reduce to a minimum intellectual re-analysis of information. It is also essential, however, that the product of intellectual analysis be easily utilized to satisfy a wide variety of "customer" needs through timely, current services and through an organized store of readily manipulable information for whatever future uses might develop. The ability of computing systems to store information reliably and to reorganize at high speed by stored instructions provides the capability necessary to meet these objectives. In order for the store of information to continue to develop in a manner that helps assure its continued usability in the face of changing user needs, utilization of a computer composition system of the type described here is essential. To that end, the computer composition of all CAS publications is anticipated within three years.

BACKGROUND

Several techniques are currently used to compose CAS publications. *CA* issues are composed through hot-metal typesetting by the Monotype process. Offset printing plates are produced from signature setups of the metal type. The nearly five million characters required to typeset one issue of *CA* are drawn from a set of approximately 1500 different characters (*2*) when case, light- or boldface, size, and superior-inferior variations are considered. Prior to 1959 all CAS publications were composed by hot-metal typesetting. Table II illustrates the major changes in CAS typesetting since then. Since 1959 not only has the number of CAS publications increased, but a drastic change in the techniques of composition has also occurred, to the extent that in 1965 approximately 35,000 of 71,282 pages of CAS publications were composed by methods other than hot metal processes.

The first CAS publication to be converted to cold-type composition was the *CA* volume Formula Index in 1959. This index is composed from Formula Index manuscript by, first, varityping entry information onto cards, one line per card, and second, photographing the line entries on the varityped cards with line-at-a-time, sequential card cameras to produce the offset negative. Since 1962 the *CA* volume Author Index and since 1964 the *CA* volume Subject Index have also been composed by this process.

The first fully computer-based CAS publication was *Chemical Titles (CT)*, the first issue of which appeared in 1961 (*3*). A new dimension of chemical information handling was created by this CAS application of the KWIC concept developed by Luhn. Not only was the computer utilized to sort, arrange, permute, correlate, and print with great speed, but it also became possible to store the information for reprocessing and custom searching.

In the computer preparation of *CT*, the article title and bibliographic information are keypunched into cards and transferred to magnetic tapes (*4*). The titles are ordered against a list of words prevented from indexing—*i.e.*, a stop list—to produce the permuted Keyword-In-Context Index along with the bibliography and author index sections of each *CT* issue. The printed copy is optically reduced to the proper page size for offset negative production.

In 1965, CAS began its first abstract journal produced with the aid of the computer, *Chemical-Biological Activities (CBAC)*. *CBAC* can be composed in upper and lower case with superscript and subscript characters through use of a specially designed 120-character computer print chain. The indexing density, and thus the retrieval capability of *CBAC* magnetic tapes, is superior to that of *CT*

Table II. CAS Typesetting

Prior to 1959	Hot metal composition
1959	Varitype-Fotolist Volume Formula Index
1961	Computer-based *Chemical Titles*
1962	Varitype-Fotolist Volume Author Index
1964	Varitype-Fotolist Volume Subject Index
1965	Computer-based *Chemical-Biological Activities*
1967	Computer-based Formula Index Computer-based Issue Indexes Computer-based Volume Author Index
1968	Computer-based Subject Index
1969	Computer-based *CA* issues
1966–1969	Additional computer-based services

tapes because index terms from the article digests are available in addition to index terms from the titles.

At present each complete *CA* issue contains four computer-based indexes that serve that particular issue. They are the Keyword, Author, Numerical Patent, and Patent Concordance Index. Each issue of each of the five *CA* section groupings (*5*) contains the complete *CA* Keyword Index for the full issue from which the grouping was taken. In order to provide timely issue indexes, speed is essential in current issue index composition. The necessary speed is readily obtained by keypunching the entries into cards at the time of abstract issue composition, computer sorting the entries, and computer printing the copy in a form suitable for preparation of offset printing plates. The index entries are also stored on magnetic tape for future uses. This procedure has been followed for three years and the tapes have been accumulated.

The conversion of all CAS services to a computer base is scheduled to be complete by 1970.

OBJECTIVES OF COMPUTER-BASED COMPOSITION AT CAS

The benefits of computer composition are considerably greater than would be achieved if the new techniques were to be considered only in substitution for existing composition processes. One objective of the CAS computer-based system is, of course, to help prepare the publications and indexes now produced by other processes. But among the added benefits beyond simply modernizing composition techniques will be reduced production cost, improved utilization of professional and clerical manpower, concurrent publication and tailored alerting capability, and information storage in a form suitable for additional computer processing. Sacrifices in page quality and content of CAS publications and services with the new techniques are not anticipated.

During computer-based composition a store of information on magnetic tape is prepared in such a manner that it is available for reorganization into new forms that may not be anticipated at the present time. One potential use of the machine language store of information is for in-depth retrospective searching. In addition, special compendia can be prepared by selecting information related to specific topics.

Selection of material being processed for inclusion in *CA* can provide the data base needed to prepare, regroup, publish, and/or search material specified by individual industrial, governmental, and academic institutions.

Another objective of the CAS computer-based system is to provide alerting services tailored to the needs of specific individuals or organizations. These services will be provided at publication time by scanning the information used to compose each publication and matching it against profiles of subscribers' interests. Full bibliographic references, or abstracts and indexes can be provided for the subscriber. Current and retrospective search services based on the mechanized store of information can be accommodated by the computer without re-analysis or transcription of the information appearing in the publication.

Another primary objective of computer-based composition is to supplement professional manpower by utilizing the computer to eliminate redundant effort wherever it is practical to do so. An example is the recording and control of bibliographic material during the assignment of articles to abstractors for *CA* and other publications and services. The abstracts returned to the computer system for processing into the publications will not repeat the bibliographic information, since that information was already made available in machine language at the time of abstractor assignment. This is but one simple example of multiple use of the information after a single analysis which will have a significant effect on professional and clerical manpower requirements in support of the preparation of CAS services.

Activity aimed at computer composition of primary journals is under way at American Chemical Society headquarters. Kuney (*6*) has described a cooperative venture directed to the eventual exchange of information between primary publications and secondary service activities. CAS expects, in time, to utilize primary journal abstracts, bibliographic information and other information such as structural formulas and chemical compound names directly in its publications as received in mechanized form from primary publication composition activities.

Several forms of assistance to CAS operations management will also be incorporated into the computer-based system. One of these is the feedback of production statistics related to primary journal productivity. Another is a variety of production control statistics. This type of information will be used increasingly in the CAS internal management information system to control over-all processes.

Finally, integrated processing and multiple use, made feasible by the computer, are essential to the maintenance of favorable cost-value ratios as the volume of information to be processed steadily increases.

INFORMATION REPRESENTATIONS

As noted previously, the composition of *CA* requires the use of 1450 different characters. Standard computer printers have traditionally had the capability of printing 48 to 56 unique characters. GRACE, developed for MEDLARS, has a character set of 226 printing characters plus 30 blank spaces of various widths and control characters with software functions (*1*). While this device marked a major step forward in combining speed with graphic arts quality in computer-driven printers, it does not offer the potential sought in the ultimate CAS system.

The CAS solution to the problem of character handling is threefold: first, reduction of the number of symbols (or different characters) used; second, development of a means for storing a wide range of specific characters; and third, a system of translating stored characters into the range of printing characters permissible in any given printing device. CAS has used these techniques for several years. For example, *CBAC* has been prepared on a regular basis since January 1965 utilizing a computer print chain containing a 120-character subset of the potential full range of characters. While computer printing is restricted to 120 characters in this application, information essential to the full range of printing characters for the future is stored, and translation routines are used to "spell out" characters not available in the 120-character set. When a satisfactory device capable of printing more than 120 characters is used in *CBAC* production, stored information will already be available to "direct" the printing of additionally available characters.

Keyboarding the broad range of characters required for the composition of *CA* can be accomplished by a variety of techniques. The technique CAS is using consists of assigning most of the keyboard codes to the lower case Roman and Greek letters, the cardinal numbers, and common special symbols. Capitalization is indicated by a symbol preceding the character or word to be capitalized, superior and inferior by forward and reverse platen indexes, and italics by underscoring. Infrequently used characters not contained on the keyboard are obtained by keying a special code followed by a mnemonic abbreviation.

Studies are underway aimed at the development of optimized keyboard arrangement, character representation conventions, and keyboarding short cuts. An example of a keyboarding short cut is computer (rather than operator) capitalization based on identification of a string of words as the title of an article.

The main advantages of the above approach are simplicity of use, production of easy-to-read hard copy, expandability to future needs, and adaptability to a variety of keyboarding devices—*e.g.*, tape-generating typewriters, keypunches, magnetic tape data recorders, and on-line terminal systems. For each CAS publication, format is consistent to the point that type style, size, and line arrangement can be controlled by information content without the use of format codes. In other words, rather than use codes that signify only "flush left" or "indent," for example, text is identified as a title, an author, a chemical compound, etc. In composing a particular publication this type of content identification functions as a format code, pertinent to that pûblication; but in storage and retrieval operations, where a format code would be useless, the content identification serves other than formatting purposes. To illustrate further, the content of abstract headings is consistent from publication to publication despite the fact that format varies. There is a single input format which is rearranged into a variety of output formats solely through codes which derive format from content.

Representation techniques for characters within the computer consist of two types: a string of characters and indicators in the form received from keyboard units and a string in a form more suitable for manipulation. The latter consists of an eight-bit representation of each unique character and a second eight-bit modifier that defines variations of the unique characters, such as light or boldface and superscript or subscript. Use of this two-part representation stored in two separate fields permits proper alpha-numeric sequencing in an efficient manner unaffected by the character variations.

The preparation of printed copy has long been a major barrier to computer-based composition because of limitations that exist in computer-driven printing equipment. Among these limitations are the lack of line justification capability, inferior print quality, and restricted ranges of characters. Computer-driven photocomposition equipment now exists that has much broader capabilities than either conventional chain printers or "traditional" photocomposing devices such as Fotosetter, Photon, and Linofilm (7). GRACE was previously cited. CAS, with IBM's assistance, is developing the application of the IBM 2280 film recorder for composition of CAS material.

The 2280 generates character images by electronically "painting" the characters on the face of a cathode ray tube. The resulting light image on the tube face is optically focused onto film, causing exposure. Any character that can be defined can be displayed on the equipment. Commonly available cathode-ray display equipment has 1024 addressable positions on the horizontal and 1024 on the vertical axis. The 2280 permits addressing 4096 positions on each axis. Since four times as many positions can be addressed in the same area, and the unit generates a thin beam, the resolution attained is four times as precise as that attainable with conventional units. Capacity of the standard IBM 2280 system is 200 to 300 graphic arts-quality characters per second. Modifications of the system being investigated could lead to 10 times that throughput (2000 to 3000 characters per second). In a proofcopy mode (lower resolution), even higher rates can be attained. The next phase of CAS development in this area will be experimentation with generation of chemical structural formulas utilizing the IBM 2250. This device is also a character generator, but it does not generate graphic arts-quality characters and does not print on film as does the 2280.

The formatting of pages is tailored for each CAS published service. Because the CAS system separates input techniques from storage conventions and applies standards for both, equipment and technique changes may be made without affecting the flow of material. In addition, information services can be added or modified independently of data flow or of computing, keyboarding, and composing equipment peculiarities.

PROCESS PHASES

Figure 1 illustrates the major information handling phases of the CAS computer-based system. At the head of the stream of CAS activities are the processes of journal acquisition, selection of specific papers for inclusion in CAS publications and services, and assignment of the selected papers to selected abstractors. During these processes all bibliographic information required for inclusion in the abstracts sections of *CA* and of the specialized publications and services, in the *CA* issue Author Indexes, in the *CA* volume Author Indexes, and in *CT* is prepared during one analysis. A single keyboarding, editing, and proofing of the information serves the need of all subsequent uses of the information, both within the computer composition portion of the system and in record keeping required in the CAS author records and control activities.

Following acquisition and assignment, the original article is reviewed in depth and digests or abstracts are prepared. During the analysis of the article and the editing of its abstract, entries for the *CA* issue Subject Index, Patent Concordance information, and information retrieval role indicators are prepared. If the given article is intended for a specialized publication as well as for *CA*, the appropriate digest is also generated at this step. Computer programs direct the appropriate portions of the analysis to the proper points within subsequent processes and, when necessary, merge the previously recorded bibliogra-

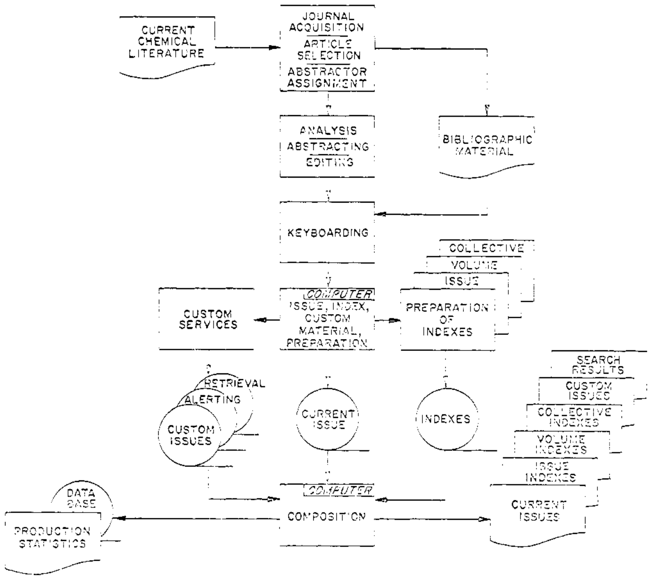

Figure 1. CAS integrated information handling.

phic data with the abstract body for final publication. This computer directed processing is also applicable to author indexes. Although the volume Author Index contains much more information than does the issue Author Index, both are derived by computer program from data present in the original analysis.

The formats of CAS publications and services are designed to meet the needs of the individual who uses the information. Original recording methods, however, must take into account convenience to the professional analyst, convenience and accuracy of recording by keyboarding personnel, and information identification required for automatic selection by the computing system in the preparation of special indexes. After information has been recorded in the computer system, it is rearranged and printed in a form convenient for editing. Provisions to facilitate the handling of editorial changes are provided by the computer. Correction entries are applied to single lines, words within lines, and in some cases to specific characters, without regard to other portions of the information. Based on the context of the information, diagnostic comments concerning completeness and the validity of checkable fields are listed as an aid to the editing process. Finally, the edited versions of the information are composed into formats suitable for the user of the publication.

During the composition process the information required for indexes are selected by program and recorded onto other information files. Examples are keywords, author names, keyword-in-context lines, Patent Concordance cross references, periodicals covered, and molecular formulas. Each of these files is in turn sequenced and formatted, and master copies are composed. *CA* issue indexes prepared in this manner can be cumulated from issue to issue and merged at the end of a cycle for the purpose of preparing the volume indexes. Additional information except for such modifications as error corrections and terminology changes need not be entered into the system in order to prepare the volume indexes. In current practice only the issue patent indexes are cumulated for volume indexes, but the other indexes will gradually be converted to greater dependence on computer techniques. The volume index information will similarly be cumulated for subsequent merging and preparation of collective indexes. As with the issue-to-volume cumulation, the information required for the preparation of the collective index will not require either professional analysis or clerical keyboarding and re-entry to the system except for error corrections and such modifications as nomenclature changes for chemical compounds.

Individualized alerting services can be provided by comparing subscribers' interest profiles. CAS now offers two such services, *CT* Search and *CBAC* Search, through which subscribers may use their own or CAS computers to obtain biweekly matches of issue contents and profiles. Magnetic tapes used for issue preparation are retained. They represent an integrated mechanized store of information available for subsequent uses.

CAS production statistics and statistics related to the yield of primary journals can also be prepared as a by-product of the computer-based composition process.

LITERATURE CITED

(1) Austin, C. J., "Experience with the GRACE System," pp. 61-70, in "Automation and Electronics in Publishing," L. H. Hattery and G. P. Bush, Eds., Spartan Books, New York, 1965.

(2) The word "character" has two definitions in this paper. As used in this statement A, a, *a*, etc., are "different characters." From the CAS computer storage standpoint, however, these three letters are considered variations in form of the unique character "a." Therefore, while 1450 "different characters" are required to compose *CA* issues, the CAS coding technique treats these 1450 as a much smaller set of "unique characters," the variations in the computer record being defined by modification flags (bits).

(3) Freeman, R. P., Dyson, G. M., J. CHEM. DOC. **3**, 16 (1963).

(4) In mid-1966, *CT* input was converted to keyboarding directly onto magnetic tape.

(5) Each *CA* issue contains 74 sections, analogous to chapters, each dealing with a chemical or chemical engineering topic. Five groups of sections—*i.e.*, "section groupings," are also published as separate units. The five groupings consist of sections 1-15, 16-30, 31-44, 45-55, and 56-74.

(6) Kuney, J. H., Lazorchak, B. G., Walcavich, S. W., J. CHEM. DOC. **6**, 1 (1966).

(7) Anderson, P. L., "Compatibility of Input and Output Devices," pp. 91-101 in "Automation and Electronics in Publishing," L. H. Hattery and G. P. Bush, Eds., Spartan Books, New York, 1965.