Design and Operation of a Computer Search Center for Chemical Information*

MARTHA E. WILLIAMS and PETER B. SCHIPMA
IIT Research Institute, 10 West 35th St., Chicago, Illinois 60616

Received January 19, 1970

The objective of the Computer Search Center (CSC) of the Information Sciences section of IIT Research Institute (IITRI) is to provide a link between a wide variety of users and the rapidly expanding information resources in machine-readable form. Because none of the available computer search programs met the criteria of the center, and because of the need to handle a variety of data bases, new general purpose computer programs were written, and a tape format was developed so that a wide variety of data bases can be searched by the same computer program. The center was designed to provide current awareness and retrospective search services from both document-type and data-type computerized data files. The desire to develop transferable programs for use at many installations prompted the adoption of the machine-independent compiler language PL/1 and the use of IBM 360 series computers. The objective of education and training led to the development of a "Search Manual" for profile preparation, the development of a workbook in "Modern Techniques in Chemical Information," the teaching of a new academic course, and the presentation of seminars.

There is a large volume of chemical information that now exists in machine-readable form, and there are many chemists who are potential users of this information. A potential market exists, but there is a real problem in devising good methods of bringing the users to the information sources or disseminating the information from these sources to users.

In July 1968, the Information Sciences section of IITRI began development of a Computer Search Center (CSC) for handling and disseminating information from machine-readable data bases. In addition to creation of an operational computer storage, search, retrieval, and dissemination system, IITRI has instituted educational and training programs, the purpose being not only to develop a center, but to ensure its use.

DESIGN OF COMPUTER SEARCH CENTER SYSTEM

The Computer Search Center was designed with the very general objectives of becoming a one-stop information center for handling numerous machine-readable data bases and meeting user needs by providing the desired sources and services with minimal restrictions and a high degree of flexibility. These objectives, together with the practicality of financial limitations, led to the establishment of design requirements and the development of special features for the CSC system. Requirements included program transferability, machine-independence and installation-independence, ability to handle numerous data bases, development of general purpose programs, and modularity. Special features include aggregation of profile terms; left as well as right truncation of profile terms; free form Boolean logic; removal of redundant searchable terms; options for sorting of output; options for media on which output is printed; and designation of hit terms, index terms, and weight for each output citation.

One of CSC's objectives was the development of programs that could run at a variety of installations. Inasmuch as the IBM 360 family of computers represents a large segment of the computer field, we decided to program for the 360. Programs were written to be run on 360's from the Model 40 on up. They require a minimum of two tape drives, one or more disks and, assuming approximately 3000 search terms (200 profiles of 15 terms each), 256K bytes of core storage.

In an effort to achieve machine independence, installation independence, and program transferability, the high level compiler language PL/1 was adopted in preference to the more economical Basic Assembly Language. The fact that this objective has been met is demonstrated by the fact that between April and September 1969 CSC programs, in both source and object code, were run at seven different computer installations on 360 Models 40, 50, 65, 67, and 75, using 2311 and 2314 disc drives with PCP, MFT, and MVT processors, under two versions of OS, 15-16, and 17. In no case was any significant problem encountered. PL/1 has proved quite satisfactory because of the facilities it provides for manipulating bit and character strings, handling multi-dimensional arrays and structures, and performing INPUT/OUTPUT operations. IITRI programmers found PL/1 easy to learn and the time required for program development was considerably less than it would have been if an assembly language had been used.

CSC programs were initially written and debugged using the RUSH (Remote-User-Shared-Hardware) interactive programming system. Using a terminal at IITRI, pro-

* Presented in Symposium on Who Reads the Chemical Literature and for What Purposes, Division of Chemical Literature, 158th Meeting, ACS, New York, September 8, 1969.

grams were written, compiled, and debugged on a 360/50 in Palo Alto, Calif. RUSH is a dialect of PL/1, and programs were developed avoiding those features and statements in RUSH that were not currently in PL/1. The transition from RUSH to PL/1 went very smoothly.

PROGRAMS

The programs were written in a modular fashion so that changes, additions, and deletions could be readily accommodated. A separate block was written for each separate operation within a program. The basic functions provided by the programs are source tape format conversion, profile preparation, search, output generation, and maintenance of statistics.

Format Conversion. Tapes are received at CSC and converted from the suppliers' to the standard IITRI format which allows an open-ended number of data elements as found on tapes from a wide spectrum of suppliers. In principle, the IITRI format is very similar to the Library of Congress' MARC II format, the COSATI Interchange Format, and the CAS Standard Distribution Format in that data elements are specified and identified by type and field length. The format conversion program checks for errors and omissions on supplier tapes, removes redundant terms, and rearranges the data into IITRI standard format. If an error is found, it is recorded for reference, but no changes to content are made. The program associates each data element with an IITRI data type number so that the search program can search for data items by type. The IITRI standard format employs a directory and a string for each citation. The string is a continuous string of characters representing all of the data for a given citation or record. The directory indicates the length of each data type and points to the location in the string where each of the data types begins. CSC programs currently allow for 99 data types. The 11 in use at present are given in Table I.

A different program is written to format each of the different tape services offered by suppliers. At present we have completed the programs for CAS Condensates, CBAC, POST and BA Previews, ISI's source tapes, and Engineering Index's COMPENDEX. Additional programs will be written for tapes from PANDEX, the Clearinghouse for Scientific and Technical Information, American Institute of Physics, and others. The decision as to which tapes will actually be run in our SDI service will depend on the results of a market survey.

The removal of redundant terms is done for purposes of economy since the machine time required to search a data tape is directly proportional to the size of the searchable file. Redundant terms are terms that appear more than once in a given citation or record. For example, if the same term appears in the title and the index terms or keywords associated with a citation, we would retain it only in the title. If a multi-word term appears in more than one permuted order, we retain it in only one order. This effects a saving of computer time because in one tape there is as much as 75% redundancy in terms. Redundancy removal is performed by a subroutine which can be used or omitted depending on the data base to be formatted.

Table I. List of Data Elements

Data Type Code	Data Element
01	Source information (coden, journal reference, pagination and dates)
02	Title of article
03	Author(s)
04	Short journal title
05	Keyword(s)
06	Registry number
07	Molecular formula
08	Corporate author
09	Abstract or text
10	BA cross code
11	BA biosystematic code

Profile Preparation. The profile preparation programs are called DKEDIT (DecK EDIT) and INPUTR (INPUT-eR). DKEDIT checks for keypunch and syntactical errors in the profiles, and it maintains counters to check such things as the equality of the number of term punched cards in a profile and the number of terms designated by the profile header card. When errors are detected they are flagged, and specific error messages are printed out indicating the types of errors present.

INPUTR aggregates the terms contained in all of the profiles that are to be run. Associated with the unique terms, by an algorithm, are numbers for all profiles in which any term occurred. Since the number of profile terms is one of the most significant factors affecting search time, aggregation provides economic advantages. Aggregation reduces the number of search terms from 15 to 23% depending on the total number of profile terms and the similarity of the profiles.

INPUTR contains the subroutine EORP (Early Operator Reverse Polish) which rearranges free form Boolean logic expressions into unambiguous parenthesis-free Polish notation. INPUTR also processes the link and weight parameters and provides all profile information to the search and output preparation programs.

Search. There are two inputs to the search program, the profiles that have been processed by DKEDIT and INPUTR, and the data tapes that have been processed by FORCON. The search program, SEARCH, sequentially matches each citation against the aggregated list of profile terms. Each profile term is designated by term type (author term, text term appearing in a title or as a keyword index term, CODEN, corporate author, etc.) and is matched against that portion of the citation or record containing the appropriate type of data. For example, an author term in a profile is matched against the author portion of the citation, and if no author is given in the citation, no further scanning of that citation against that profile term takes place.

Output. Citations that are hits in the search are read onto a disk file. The FORMAT program formats the citations as they appear on the printed output. The for-

Table II. Term Aggregation Reduction Ratios

Profile Terms before Aggregation	Search Terms after Aggregation	% Reduction
800	674	15.7
971	791	18.5
2849	2243	21.3
2931	2275	23.0
3321	2729	17.8

matted citations are sorted by the SORT program and printed by the OCP (Output Copy Preparation) program. Citations are sorted according to user specification, by reference number, author, or weight, and printed on 5 × 8 cards or paper as requested.

The statistics program maintains a variety of statistics needed for both acounting and research. A flow chart indicating program interfaces is given in Figure 1.

ANALYSIS OF MULTIPLE DATA BASES

Initially, we are providing a regularly scheduled SDI service from the CAS Condensates tapes. However, our system and programs have been designed to offer both SDI and retrospective services from virtually any of the document type data bases. This being the case, format conversion is a very important step in the system since format and contents of data tapes vary considerably within and between suppliers. This is evident from the results of a detailed analysis of six different source tapes. The source tapes studied were:

Source	Data Tape
Chemical Abstracts Service	CBAC Vol. 8, issue 13, 1969
Chemical Abstracts Service	POST-J Vol. 4, issue 2, 1969
Chemical Abstracts Service	COND Vol. 70, issue 13, 1969
Biological Abstracts	BA Previews Vol. 50, issue 9, 1969
Engineering Index	COMPENDEX Test tape, April 1969
Institute for Scientific Information	ISI Source Test Tape, 1969

It must be realized that these are individual tape issues, and data conventions noted from these tapes may not correspond to the current conventions of a supplier or to conventions noted by other tape users who have analyzed different issues of the same tape service.

The analysis has indicated a real need for standardization among source tapes and for documentation regarding the tapes.

All issues were available on 9-track tape with a recording density of 800 bpi. From this point on, the areas of uniformity begin to disappear.

The variety of physical tape formats included variable length, fixed length, and variable number of fixed-length field records. There is no uniformity in data elements nor in the contents of the elements. For example, some services included source data (pagination, volume, issue, year) with the CODEN, others with an abbreviated journal name, and others in separate data elements. We defined 16 distinct types of data found on the six tapes studied, but no tape contained more than half of the 16. The methods of representing similar data types also vary. For example, author's names and initials are represented in at least four variations on the six tapes.

This lack of consistency between tape sources is taken care of at CSC by writing a formatting program for each tape. Internal lack of consistency within a single tape, or from one tape to another of the same data source, is a very present and serious problem and requires many error checking routines in a given program. Both types of inconsistency should be decreased by tighter quality control at each source and by cooperation between suppliers.

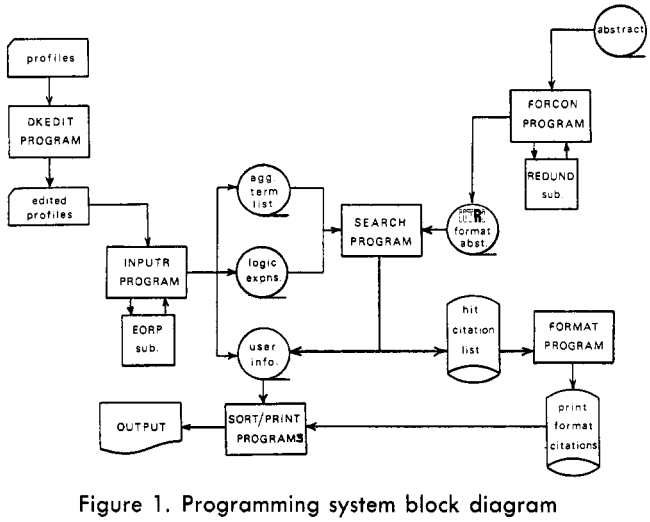

Figure 1. Programming system block diagram

PROFILE OPTIONS

The CSC system is flexible with regard to profile preparation, search capabilities, and output options.

Logic. In preparing his profile, a user can include up to 100 terms. He can relate terms and/or groups of terms to one another by means of the standard Boolean operators AND, OR, and NOT. The terms and operators are used in free form logic expressions to express the specific interests of the user.

Links. Terms that are semantically associated can be linked together in a single expression. That is, several terms that are synonymous, related, or hierarchically broader and narrower, can be represented by a single alphabetic character. This simplifies the user's task of writing his logic expression. He can merely specify a link designator rather than indicate the multiple terms joined by the link in cases where any one of the terms would be equally satisfactory in the logic expression. For example, if a user were interested in halogens and oxygen, he could use the link designator A to represent the terms halogen - fluorine - chlorine - bromine - iodine. In writing his expression, he would specify (A&006), where 006 is the term number of oxygen.

Truncation. Since many data bases include titles, which are author-generated and therefore uncontrolled, it is necessary to include in one's profile all forms of a desired term to ensure retrieval of the desired information. To simplify this task of specifying all possible relevant word forms and fragments, CSC has allowed all options in truncation. Left, right, both, and none modes of truncation are permitted. The singular and plural forms of a term can be retrieved by using right truncation—e.g., chloride[a] ([a]denotes truncation) would retrieve both chloride and chlorides. An example using the term AZO is given in Figure 2.

Truncation has been employed by all of the participants in the CSC SDI program. Considering all the profile terms in several runs:

No truncation was used for 46% of the terms
Left truncation was used for 5% of the terms
Right truncation was used for 36% of the terms
Both truncations were used for 13% of the terms

Weights. CSC profiles permit the assignment of weights by users to further refine their profiles. Weights are numer-

Mode	Input Format	Action	Examples
0	AZO	Retrieves only the term AZO	AZO
1	°AZO	Retrieves any term ending in AZO	AZO, DIAZO, HYDRAZO
2	AZO°	Retrieves any term beginning with AZO	AZO, AZOXY, AZOLE
3	°AZO°	Retrieves any term in which AZO appears	AZO, DIAZO, HYDRAZO, AZOXY, DIAZOMETHANE

° Denotes truncation

Figure 2. Modes of term truncation

ical values from 0 to 9 assigned to terms to specify their relative importance. If a user employs weights in his profile the output is arranged in descending weighted order so that those citations with the highest weights—presumably the references that are of most significance to the user—will be on top.

Searchable Terms. With respect to searching, the CSC system is flexible in that a user can search not only on text-type terms, with all their associated parameters, but also on any of the term types included on the tape. These are CODEN, title, author(s), short journal title, keyword(s), registry number, molecular formula, corporate author, abstract, BA CROSS code, and BA Biosystematic code. While the system can be expanded to include any term type, at present searches are being made of the CAS Condensates tape, and the searchable items on that tape are CODEN, title, author(s), short journal title, keyword(s), and corporate author.

Output Sort. With respect to output, several options are open to the user. He can specify an upper limit to the number of hit citations he would like to have printed out for him. He can indicate whether he wants his output printed on cards or paper, and he can specify the way in which the output should be sorted—alphabetically by first author's last name, numerically in ascending order by reference number, or in descending weight order.

OUTPUT SENT TO USERS

Output sent to users is of two types, header information and citations. The header card as shown in Figure 3 indicates the user profile number, the tape service and issue of the tape that was searched, the number of citations that were on tape, the number of citations that were hit citations for the user's profile, the number of citations that were printed, and the date of the search. The header card is followed by the hit citation cards.

DECEMBER 31, 1969

PROFILE C1A0100628

CA VOL. 71, NO. 23 WAS SEARCHED

ISSUE CONTAINED 4301 CITATIONS

HITS FOR THIS ISSUE: 16

NUMBER OF HITS PRINTED: 16

COMPUTER SEARCH CENTER
IIT RESEARCH INSTITUTE
10 WEST 35TH STREET
CHICAGO, ILLINOIS 60616
312/225-9630

Figure 3. The header card

There is one 5 × 8 inch card for each hit (Figure 4). A citation card includes: abstract number, tape source, volume and issue number, profile number, authors (as many as are given on the source tape), corporate authors, full title, primary source information (journal, volume, issue, date, and pages), CODEN, index terms that were included on the source tape, search terms present—i.e., those profile terms that were hit terms for the particular citation and weight for the citation.

EDUCATION AND TRAINING

IITRI has followed several routes for providing education and training to current and potential users of machine-readable data bases. These include seminars, formal university courses, short courses, development of a workbook for "Modern Techniques in Chemical Information," and preparation of a "Search Manual" to guide users in preparing profiles for searching computer-based information files.

Course in Modern Techniques in Chemical Information. One of the more significant education efforts has been carried out in cooperation with Illinois Institute of Technology, the university with which IITRI is affiliated. During the 1969 spring semester a new course was offered at IIT, "Modern Techniques in Chemical Information." The course was made available to second year graduate and upper division undergraduate students in the Chemistry Department. This course replaced the traditional chemical literature course, and the chemistry graduate students were given the option of taking the Modern Techniques course in lieu of a second foreign language. One hundred per cent of the graduate students opted for the course. Members of the IIT staff who serve on graduate advisory committees willingly accepted this change as a significant improvement in the formal training for the Ph.D degree. The course was made available through a subcontract from the IITRI Computer Search Center program to the Chemistry Department at IIT, and the course was taught by Paul E. Fanta of IIT and Martha E. Williams of IITRI.

The course covered techniques of storage, search, and retrieval of chemical information. Specifically, it stressed the fact that chemical information exists in many different forms, both printed and machine-readable, and if the chemist is to make good use of the multiplicity of available data bases and collections he must expand his horizons and be prepared to use the computerized files as well as the traditional collections. Information resources and methods of retrieval were considered from the viewpoint

ABSTRACT NO. 109992 CA VOL. 71, NO. 23 PROFILE C1A0100628

HODGSON B, WALKER JS. (UNIV. MANCHESTER INST. SCI. TECHNOL., MANCHESTER, ENGL).

TYROCIDINE-SYNTHESIZING CELL-FREE EXTRACT FROM BACILLUS BREVIS A.T.C.C. 10068.

BIOCHEM. J. VOL. 114, NO. 1 12 PP., 1969. (ASTM CODEN: BIJOA)

INDEX TERMS: PEPTIDES SYNTHESIS ANTIBIOTICS TYROCIDINE

SEARCH TERMS PRESENT: PEPTID SYNTH

WEIGHT FOR THIS CITATION: 0

Figure 4. The citation card

of information systems, and the general problem was considered to be the retrieval of specific data from a data store.

Inasmuch as none of the available chemical literature textbooks provide adequate coverage of the modern techniques and sources of chemical information, staff members from both IITRI and IIT (Eugene S. Schwartz and Martha E. Williams of IITRI and Paul Fanta of IIT) developed a syllabus and workbook for the course, "Modern Techniques." The objectives and contents of the book are described in the following section.

In addition to acquainting the student with the traditional and modern methods of handling information and sources of information, each of the students participated in an SDI program. Instruction in profile preparation was provided both through lectures and through study of the "Search Manual." Students became acquainted with the problems and techniques associated with development of interest profiles including selection of terms, truncation of term fragments, development of expressions for proper logical association of terms, use of links for grouping terms within an expression, and assignment of weights.

Students conducted manual searches of an issue of *Chemical Abstracts* in two subject areas—one organic and the other inorganic. After completing the manual searches, they prepared interest profiles for their two subject areas. The interest profiles were then used by IITRI in a search of the corresponding issues of the CAS Condensates tapes. Output from the SDI run was returned to the students for comparison with output from their manual searches. The time savings by the machine was dramatic and impressed students who had had to spend considerable time in conducting the manual searches.

One of the objectives was to make the students sufficiently aware of the capabilities of computer services so that when they enter the industrial community, they will request such services. These students will be the future chemists and users of computerized chemical information systems. Hopefully, in much the same way that students who use modern analytical equipment in their university laboratories demand modern equipment in the industrial laboratories that hire them, so students familiar with automated information handling will require these services from their employers.

Search Manual. In preparation for the academic program, training seminars and, independent study, IITRI developed a "Search Manual." The manual was designed to assist CSC users in developing individualized search profiles for use with CSC computer programs. In preparing a profile, the user prepares the detailed specifications he requires for retrieving citations from a data base. The manual explains the problems and techniques associated with development of search profiles. Problem areas include the inflexibility of machinable data bases; the variety of word forms (grammatic, semantic, syntactic, and generic); the variety of conventions employed for abbreviations, symbols, acronyms; the varied practices, degrees of specification, and presence or absence of controls employed in indexing and classification; and the variety of nomenclature which is a particular problem in the field of chemistry.

The special techniques of profile preparation are determination of search terms, including synonyms, higher and lower generic terms, and related terms; determination of searchable entries other than subject terms, such as authors; the use of left and right truncation for retrieval based on term fragments and distinctive letter combinations; the use of links for grouping of related terms within a logic expression; development of free form logic expressions employing the Boolean operators AND, OR, and NOT; and the assignment of weights to profile terms in accordance with relative importance of terms to the user.

Workbook for Modern Techniques in Chemical Information. The absence of any textbook providing adequate coverage of the modern techniques for search and retrieval of chemical information and of the newer—principally machine-readable—sources of chemical information prompted IITRI's development of a workbook entitled "Modern Techniques in Chemical Information."

The book was designed for use by chemists and does not require a background in computer technology, programming, or information science. It exposes the student to the potentials and limitations of information systems and sources and explains the storage, search, retrieval, and dissemination functions that characterize information systems.

The chapters or principal topics are: (1) "Information Systems," (2) "Indexing and Classification," (3) "Primary Information Sources in Literature," (4) "Patents," (5) "Secondary Information Sources in Literature-1: Abstracting Periodicals, Review Serials," (6) "Secondary Information Sources in Literature-2: Reference Works," (7) "Chemical Information Centers" including the computer searchable data bases and computer centers, (8) "Chemical Structures in Literature and Machine," (9) "Search Systems" including an introduction to computer components, programming languages, programming, and computer systems, and (10) "Information Retrieval in a Current Awareness System."

The workbook was tested via the combined IIT course "Chemistry 351" and "Chemistry 651." It is currently being reviewed through the National Science Foundations' Chemical Information Unit. After review and revision have been completed, the book will be published and distributed.

Training Seminars and Short Courses. Seminars in profile preparation and the use of the CSC are provided for participants in the SDI program. Seminars are from two to four hours in duration depending on the background or experience of attendees. Attendees are informed of the objectives and both current and planned services of CSC. The training session covers the subject contents of the "Search Manual;" the use of aids such as thesauruses and work lists; and each attendee prepares one or more profiles under the supervision of an IITRI information specialist.

Academic programs, seminars, and short courses are means of acquainting chemists with the newer information sources and services. To a certain extent we know who uses the chemical information and we know what information sources they use. We do not know very much about who will use the new machine-readable information sources and computer services. We can, however, affect their use by educating and training potential users.