Development of a Computerized Current Awareness Service using *Chemical Abstracts* Condensates*

ANITA B. ROBERTS, IEVA O. HARTWELL,** RICHARD W. COUNTS, and ROBERTA A. DAVILA
Aerospace Research Applications Center,
Poplars Research and Conference Center,
Indiana University, Bloomington, Ind. 47401

Received July 17, 1972

The experiences in developing current awareness services for selective dissemination of information from the *Chemical Abstracts* Condensates data base are described. File standardization, the weighted-term method of searching, and the algorithm used to perform the search on the CDC 6600 computer are discussed. The results of a user survey conducted in the Chemistry Department of Indiana University on the viability of the system are reported.

ARAC, Aerospace Research Applications Center, founded in 1962 as a joint effort by NASA, Indiana University, and private industrial firms, is a nonprofit information center. It is now a division of the School of Public and Environmental Affairs of Indiana University.

Prior to 1965, only the NASA data base was searched by ARAC. Computerized searches were carried out on an IBM 709 vacuum tube computer, utilizing Boolean logic on an inverted file structure. In 1965, the NASA tape format was changed to a sequential file structure. NASA's change, coupled with the replacement by Indiana University of the IBM computer with a CDC 3600, made that time appropriate for major changes in ARAC's search structure.

Primarily for economic reasons, as well as for versatility in adapting to other data bases, a Standardized File Format into which any newly acquired data base could be arranged was developed. Subsequently, other information centers have also developed their own standardized file formats to handle the data bases they process.[1]

The format, therefore, was designed to include only those data base elements expected to be common to every information file. The restructuring of the file format was coordinated with a comparative examination of four separate systems of logic for structuring search strategies.[2] The main characteristics sought were ease of learning and using a given system, its flexibility and economics. The weighted-term logic was selected as being the most versatile and economical of those compared and now forms the basis for all ARAC search strategies currently used on CA Condensates, Ei COMPENDEX, and NASA STIMS tapes.

DEVELOPMENT

The development of a chemical information system based on CA Condensates tapes was begun in 1968. Although Indiana University was unable to secure outside funding for the project, the initial development was nonetheless begun under the sponsorship of the Chemistry Department in cooperation with the Research Computing Center and the staff and facilities of ARAC. The approach taken was to reformat the CA Condensates tapes to the ARAC Standard File Format that had already been developed for use with the NASA and Ei CITE files. The obvious factor here was the economy in using existing computer software with only minor adaptations.

In this respect, the development of ARAC's system has been unique. The costly development of a separate software system was eliminated and the emphasis was placed, instead, on the interface between the computer and the user. Several chemists with graduate degrees are part of ARAC's chemical information staff and are not only familiar with the computer aspects of the CA system but can also understand the needs and problems of the research scientists. Although users are provided with a detailed instructional Search Manual, the man-computer gap is so great that the unassisted development of a satisfactory profile can be a lengthy "trial and error" process. The members of ARAC's chemical staff have been able to reduce substantially the development time by interfacing with the user regarding his initial strategy and then carefully examining the output of the first few issues of a new profile for necessary modifications. In this manner, the experience of the staff is passed on to the user, enabling him to realize quickly the fullest potential of his profile.

OPERATION

ARAC's software was originally developed for the CDC 3600 computer and has now been adapted to the CDC 6600 computer. The tape as it is received from Chemical Abstracts Service is first reformatted into two files: the Standard File Format (sometimes referred to as the linear file) and the Title File (Figure 1). During a computer run, the contents of the Standard File Format are compared with those of the profile tape, resulting in a numerical listing of the accessions pulled by each profile and a numerical listing of all unique hits for the run. The listing of unique hits is used to extract the bibliographic information for those hits from the Title File. Comparison of the numerical distribution of each profile's hits with the Subset Title File results in the final printout of the custom profiles. A compilation of total hits is also maintained by the computer for royalty payments.

*Presented before the Division of Chemical Literature, 163rd Meeting, ACS, Boston, Mass., April 12, 1972.

**To whom correspondence should be addressed.

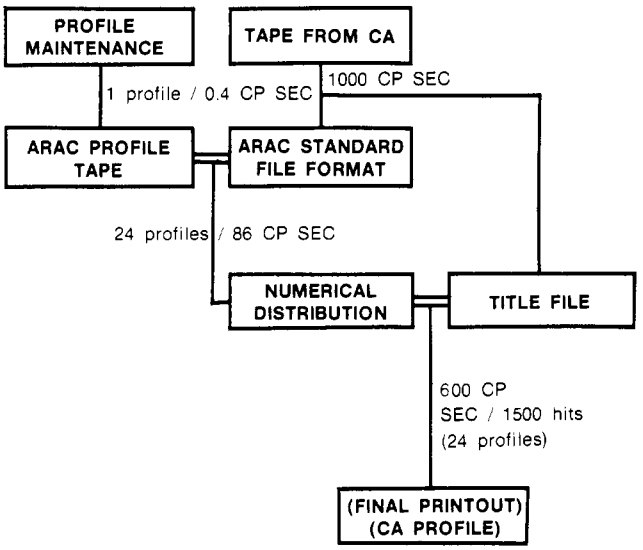

Figure 1. Processing of the CA Condensates tape

IU 001	9	299	J.A. BARTLETT	ISOTOPE EFFECTS
IU 001	9 DEUTERIUM*	9 TRITIUM*		6 DEHYDROGENASE*
IU 001	5 ISOTOPE	4 EFFECT*		3 ALCOHOL*
IU 001	3 MECHANISM	6 15		3 NITROGEN
IU 001	−9 YEAST*	9 SHINER VJ		−9 JACSAT

* = POINT OF RIGHT TRUNCATION

Figure 2. Sample weighted-term strategy

Profiles are permanently stored on a profile tape which can be modified by use of an appropriate program. Separate profile tapes are maintained for odd and even issue searching allowing the user to maintain different strategies for odd and even issues as dictated by the subject contents of the issues.

Profile term types can include author names, journal coden, words from the title or Keyword Index, as well as numbers, with an upper limit of 99 terms per profile. A sample weighted-term profile strategy is shown in Figure 2. In this example, the cutoff number is at 299; however, a value of 399 can be used. The cutoff weight is 9, although it may be any number from 1-99. The maximum term weight of 9 results in a "one-term pull," while other importance weights are assigned corresponding to a two- or three-term cross in Boolean logic. Since all terms can interact freely, consideration must also be given to the frequency of occurrence of terms, or Posting Statistics. A computer generated listing of postings for terms in an odd and an even issue is used to check the relative frequency of terms to be used in the strategy. Right truncation of terms is available, and this is indicated by an asterisk following the term. A maximum of 20 characters is used per term. Negatively weighted terms can be employed to eliminate unwanted subject material or journals to which the user subscribes if he does not wish to see citations from them.

In an actual computer run, profiles are considered internally in groups of 24. All of the terms of 24 profiles, keyed by profile numbers, are put onto one master list, and this list is alphabetized in core. Since both the profile and the document terms are in alphabetical order, total comparison of the two tapes can be accomplished by a minimum of computer matching. A further reduction

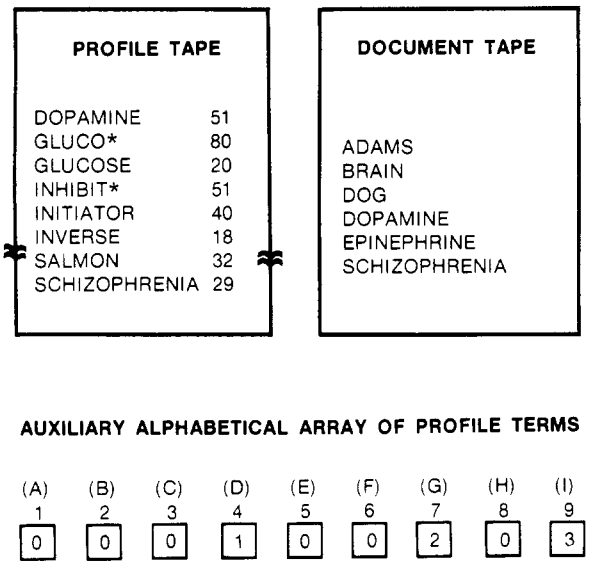

Figure 3. Excerpt from the profile tape and document tape arrays and corresponding auxiliary alphabetical array used in searching

in computer time is brought about by the generation of an auxiliary alphabetical array of profile terms, which allows the profile listing to advance by groups when matching document terms. A brief example of this process is shown in Figure 3. The auxiliary array generated from the list of profile terms indicates how many terms begin with each letter of the alphabet. In this case, considering the first profile term in the array, the computer goes directly to the first D term of a given document in the document array. Next the characters of the first computer word of the two-computer word terms are compared. Since $DOPAMINE_p > DOG_d$, the pointer to the document array advances by one. At this point, $DOPAMINE_p = DOPAMINE_d$, so the characters of the second computer word of each term are compared and found to match. The next comparison shows $G_p > D_d$, so the pointer to the document array advances by one, finds $G_p > E_d$ and advances again by one. Then $G_p < S_d$ and the auxiliary array is used to determine the advances of the pointer to the profile array by groups for each letter of the alphabet until the comparison shows $S_p = S_d$. The comparison of the first computer words is made and followed by comparison of the remaining characters. When all document terms have been checked, the next document is considered and the matching procedure is repeated. When all of the profile terms and documents have been checked, the next group of 24 profiles is considered.

The actual computer time involved in a typical run is given in Figure 1. The reformatting of the CA tape, while a major contributor to the time required, must be done only once for each issue, regardless of the number of batches of profiles run. The actual search process typically handles 24 profiles in approximately 86 cp (central processing) seconds. By far the greatest amount of time is delegated to providing the printed output which is supplied to the user. Approximately 600 cp seconds are required for an average of 1500 hits for 24 profiles. This vast amount of time used results because input/output is one of the weakest aspects of the CDC 6600 computer.

The final printout lists the pertinent bibliographic information for a citation as well as its keywords. The weight of the citation, which is an indication of the relevancy, is printed immediately to the right of the CA ab-

Table I. Prices of ARAC Services

	Paper, $	Cards, $
Alternate issues—per year	95	120
All issues—per year	155	195

Table II. Impact of Computer Service on Literature Searching Time

24%—no change
30%—decreased 0–25%
21%—decreased 25–50%
12%—increased

Table III. Effects of Computer Service on Research Efforts

42%—provided new research leads
66%—made aware of others in the field
49%—made aware work had been done
33%—made more time available for research

stract and issue number related to each citation. Output is available on either 8½ × 11 inch computer paper with an average of 3 citations per page, or on medium weight 4 × 6 inch cards with one citation per card. The charge for card service is somewhat higher than for paper because of increased mailing and card costs. Table I outlines the prices charged non-Indiana University users.

USER SURVEY

To assess the value of the service to subscribers, a survey was conducted of the Indiana University Chemistry Department users. Of 67 questionnaires sent out, 33 responses were received—17 from faculty members and 16 from graduate students. The average reported time spent on searching the literature each week was found to be 5.4 hours. The impact of the computer printouts on literature searching time is shown in Table II. Whereas 24% found their literature searching time remained unchanged, 51% felt that the time spent searching the literature decreased, in some instances by as much as 50%. Although others found that their total literature searching time increased, the benefits increased as well, as shown in Table III.

It was of interest to discover how people divide their literature searching time between various available sources (Figure 4). The ordinate of the graph represents the total hours per week devoted by the group as a whole to each particular source. Therefore, while the time spent looking at *Current Contents* or *Chemical Titles* is small when considered as a total, for the few users, these two sources represent approximately 50% and 23%, respectively, of their literature searching time. The number of persons subscribing to each of these sources, as well as to individual *Chemical Abstracts-Section Groupings*, decreased with the advent of computer searching. The greatest time is allocated to the searching of key journals and of contents pages supplied by the Chemistry Library. The time devoted to each of these also decreased significantly when the computer printouts were made available. The average user spends 23% of his literature searching time reviewing his computer printout and items announced by it, or just over 1 hour per week. In general, it appears that the computer printouts can replace similar alerting services, but they do not obviate the need for searching key journals.

Finally, when asked about their reaction to a discontinuation of the service, 57% said they would be very disappointed, 36% mildly disappointed, and 3% noncommittal. The 57% group said they would continue their subscription even if the Chemistry Department would no longer finance the service.

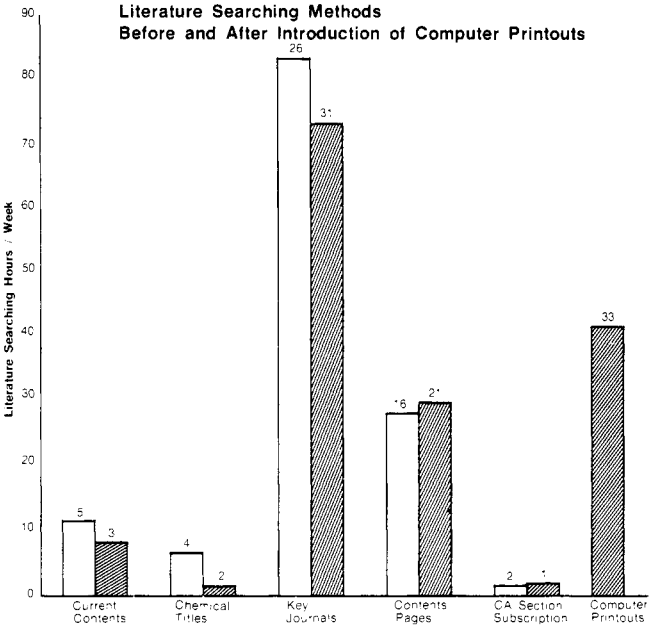

Figure 4. Relative importance of several literature sources. The numbers at the top of each bar indicate the number of users of each source. White bar = before; cross-hatched bar = after

CONCLUSIONS

Although originally developed for use by the Indiana University Chemistry Department, ARAC's CA Condensates service now has 107 subscribers to 164 profiles, covering both industrial and academic scientists. Within the IU system, the Bloomington campus, regional campuses, the Medical School, and the Dental School all subscribe to ARAC's service. The University feels that the chemical information service is an important adjunct to its library system, as well as a valuable training aid in the educational development of graduate and undergraduate students. The Chemistry Department survey, as well as unsolicited comments from other subscribers suggests that ARAC's CA Condensates service, while of simple and economical design, is a viable system and a valuable asset to its users.

ACKNOWLEDGMENT

The authors thank John M. Knego, Chemistry Librarian, for his key role in arranging the cooperative agreement between the Chemistry Department and ARAC.

LITERATURE CITED

(1) Williams, Martha E., and Schipma, Peter B., "Design and Operation of a Computer Search Center for Chemical Information," *J. Chem. Doc.* **10**, 158–62 (1970).
(2) Sprague, Ralph H., Jr., "A Comparison of Systems for Selectively Disseminating Information," Ph.D. thesis, Indiana Business Report No. 38, 1965.