Online Searching: Full Text of American Chemical Society Primary Journals

SELDON W. TERRANT, LORRIN R. GARSON,* and BARBARA E. MEYERS

Research and Development Department, Books and Journals Division, American Chemical Society, Washington, DC 20036

STANLEY M. COHEN

Research and Development Division, Chemical Abstracts Service, Columbus, Ohio 43210

Received February 27, 1984

During 1980–1982, the Books and Journals Division of the American Chemical Society conducted an experiment involving online access to the full text of its primary journals. The purpose of the experiment was to determine the feasibility of providing online access to the Society's primary journals...whether the concept would be acceptable to the scientific community and whether such a service would be economically viable. The experimental online file was accessed by several hundred volunteer participants. The reactions of the participants, determined by online searching, mail questionnaire surveys, and face-to-face discussions, are reported in this paper. Users generally found full-text searching to be a powerful method of locating information in the main text of the primary journal articles. Many of the participants indicated the need for a larger file, including non-ACS journals. The participants also indicated that a lower connect hour charge would be desirable. In June 1983 online access to 18 ACS primary journals (for the period 1980–) became available to the public.

INTRODUCTION

An experiment was conducted from 1980 through 1982 to evaluate online access to the full text of American Chemical Society (ACS) primary journals. The experiment began in 1980 with a small file, consisting of one journal (*Journal of Medicinal Chemistry*), that was established as a private database at Bibliographic Retrieval Services, Inc. (BRS). A dozen volunteers were involved in the initial experimentation. On the basis of the general acceptance of the concept by these volunteers, the file was expanded in 1981 to contain the full text of the 16 ACS journals that were available in machine-readable form at that time. Since that time, the online file at BRS has been updated every 2 weeks. The data consisting of primary journal text were derived from the computer-based journal composition system operated for the Books and Journals Division by Chemical Abstracts Service (CAS).

About 300 people agreed to evaluate this expanded online file. User reaction was sufficiently favorable to warrant continued experimentation during 1982 with a file containing about 25 000 articles from 18 ACS journals. The 1982 work involved another 250 volunteer participants. Several different methods of obtaining user reaction to the file were used. An online questionnaire was utilized in which sets of multiple choice questions were displayed at log-off time to capture the impressions of the searchers while the just-concluded online sessions were fresh in their minds. Additionally, follow-up survey questionnaires, face-to-face discussions, and telephone contacts were used to obtain feedback from the participants. On the basis of the favorable response, ACS began to offer ACS JOURNALS ONLINE (journals for the period 1980–) to the public in June 1983 via a private database at BRS. Although in the past few years the full-text files of other publications have been available online (Mead Data Central's *Lexis* and *Nexis* and portions of John Wiley and Sons' *Encyclopedia of Chemical Technology*), ACS is the first scholarly association to make its publication available in this manner.

JOURNAL COMPOSITION SYSTEM

Tapes containing the ACS primary journals' text are routinely produced by the journal composition system[1,2] that was developed by CAS in cooperation with the Books and Journals Division of the ACS. This system generates a database in which the textural components (e.g., titles, author, paragraphs, footnotes, etc.) are uniquely identified by means of specific data elements. The manner in which text components are stored on the database is the same for all the journals presently produced by the system and is independent of how the component in question is formatted on the final journal pages. This type of identification is known in the publishing industry as generic coding. The unique identification of the textual components proved to be highly valuable in that all of these components were easily and consistenly identified by BRS. Tapes sent to BRS were formatted in CAS's Standard Distribution Format.[3] The various text fields were then directly mapped into the appropriate search and display fields loaded at BRS. Data are loaded into the online file within 1 week or so after composition is completed; depending upon geographical location, users in the U.S. can access data in the online file about the same time they receive the corresponding hard-copy journals.

Data files generated by the journal composition system contain information about the size and location of the graphic components that are manually stripped into the final pages. Records containing this graphic-positioning information were removed by the programs that produced the tapes sent to BRS. While the online file contained no graphic information, the captions that describe figures were retained. This text is available both for search and for display in the online file.

SCOPE OF ONLINE FILE

During August 1982 through October 1982, 144 of the 250 individuals who volunteered participated in an experiment with an online file containing the full text of 18 ACS primary journals. The experimental results discussed in this paper pertain to the 1982 phase (essentially equivalent to the 1981 results). The coverage of the online file during this final phase of the experiment is shown in Table I. The file has been updated every 2 weeks as the primary journal text became available from the journal composition system that produces the traditional hard copy. The numbers in Table I represent the status of the databases as of November 17, 1982. During the final phase of the experiment, the database contained 25 000 documents (articles). In February 1984, the file contained 38 519 documents.

0095-2338/84/1624-0230$01.50/0

Table I. Primary Journal Experimental Database

journal	no. of documents	coverage[a]
Journal of the American Chemical Society	5017	1980–1982[b]
Journal of Organic Chemistry	3774	1980–1982
Journal of Medicinal Chemistry	2981	1976–1982
Biochemistry	2851	1980–1982
Inorganic Chemistry	2643	1980–1982
Journal of Physical Chemistry	2489	1980–1982
Analytical Chemistry	1960	1980–1982
Macromolecules	986	1980–1982
Journal of Agricultural and Food Chemistry	967	1980–1982
Journal of Chemical and Engineering Data	413	1980–1982
Industrial & Engineering Chemistry Process Design and Development	357	1980–1982
Industrial & Engineering Chemistry Product Research and Development	345	1980–1982
Organometallics	398	1982
Industrial & Engineering Chemistry Fundamentals	256	1980–1982
Journal of Chemical Information and Computer Science	188	1980–1982
Accounts of Chemical Research	175	1980–1982
Environmental Science and Technology	148	1982
Chemical Reviews	53	1980–1982
total	25901	

[a] Through November 19, 1982. [b] Coverage of *Journal of the American Chemical Society* begins in July 1980.

STRUCTURE OF ONLINE FILE

When the text of the primary journal articles was loaded at BRS, it was structured into fields for both search and display purposes. Table II describes the fields contained on the online file. All fields shown in Table II can be displayed. All fields are searchable, except the occurrence field (OC), which is generated during the search.

SEARCHING FULL-TEXT FILE

A comprehensive description of the BRS searching system can be obtained from the BRS manual[4] and from a manual[5] published by ACS for users of the primary journal full-text file. A number of searches are illustrated below. These searches are intended to illustrate system features that are particularly important in searching as online full-text file.

Boolean Operators. Our experience has shown that Boolean operators, which require coordinated search terms to be in relatively close proximity (within the same paragraph or closer) to each other in a document before the document can be retrieved as a hit, are of particular importance in full-text searching. Considering the massive number of index entries for a typical full-text journal article in an online file relative to the small number of index entries for a bibliographic file, the potential for false coordination of search terms logically linked in full-text files at the document level is apparent. Table III shows the Boolean operators in the BRS system and their functions.

```
1-:    nucleophilic and displacement and fluorine
RESULT       92

2-:    nucleophilic same displacement same fluorine
RESULT       14

3-:    nucleophilic with displacement with fluorine
RESULT        5

4-:    nucleophilic adj displacement adj fluorine
RESULT        3
```

Figure 1. Searching with different levels of search operators.

Figure 1 illustrates a search session in which three search terms were run against the ACS full-text file with, in turn, each of the four Boolean operators shown in Table III. Notice in Figure 1 that, as the search operators become more restrictive, the number of documents retrieved decreases. When document-level logic is used, the number of hits obtained is usually much larger than if a more restrictive operator is used, but the potential for false coordination is large.

For example, when document-level logic is used, a primary journal article would be retrieved if the three coordinated search terms were widely separated as in the 6th, 48th, and 80th paragraphs. Occurring this far apart in the text, the three search terms would have a high probability of being unrelated, thus probably resulting in many false drops. Contrast this to the use of sentence-level logic. Here, the three search terms must be in the same sentence before a hit is produced. In this case, the three terms would have a higher probability of being related. This is not surprising as authors use words within paragraphs (SAME), and certainly within sentences (WITH), that are related to one another.

Finding Relevant Information within Retrieved Documents. Once primary documents are obtained as hits, the searcher wishing to display those portions of the retrieved articles containing the hit search terms is presented with a substantially different situation than in the case of all of the retrieved documents; there is just too much text involved to make this practical. The average full-text record contains about 26 000 characters, as compared to 200–500 for a bibliographic record...2 orders of magnitude greater. In full-text files, a

Table II. Online Search and Display Fields

field symbol	description of field
AN	a unique accession number (manuscript number) assigned to each primary document
CD	journal CODEN
SO	source field containing bibliographic information
TI	title
AU	author(s)
AF	author affiliation—indicating the institution with which the author(s) is(are associated
AB	abstract
TX	paragraph text; each individual paragraph is contained in a separate repetition of the TX field; TX(1) contains the first paragraph of an article, TX(2) contains the second paragraph, etc.
FN	numbered footnotes; each individual footnote is contained in a separate repetition of the FN field; FN(1) contains the first footnote of an article, FN(2) contains the second footnote, etc.
RF	unnumbered references; each individual reference is contained in a separate repetition of the RF field; RF(1) contains the first reference in an article, RF(2) contains the second reference, etc.
CP	figure captions; each individual figure caption is contained in a separate repetition of the CP field. CP(1) contains the caption associated with Figure 1, CP(2) contains the caption associated with Figure 2, etc.
RN	CAS Registry Numbers; each individual Registry Number is contained in a separate repetition of the RN field; the individual RN fields are identified as RN(1), RN(2), etc.; associated with each individual Registry Number is either a name for the substance represented by the Registry Number or a number that is contained in the text of the article at the point in the text where the substance is discussed

Table III. Boolean Operators[a]

Boolean operator	function
AND	requires search terms to be in the same document
OR	requires one or more search terms to be present in the same document
NOT	requires a specified term not to be present in the document
SAME	requires search terms to be in the same paragraph or field
WITH	requires search terms to be in the same sentence
ADJ	requires search terms to be adjacent to each other in the order specified

[a] Generally, only AND, OR, and NOT are considered to be true Boolean operators whereas SAME, WITH, and ADJ are thought of as positional or linking operators. In this paper, we consider all of these to be Boolean operators.

OC	PARAGRAPH	SENTENCE	NS-WORD
TI	(1)	1	11
AB	(1)	1	25
TX	(1)	2	7
TX	(2)	2	2
TX	(8)	6	23
TX	(18)	3	5
TX	(44)	4	16
FN	(2)	1	23
FN	(8)	1	10

Figure 2. Occurrence field. NS-WORD means non-stop list words. The word count is based upon words from the text that are not contained on the BRS stop word (trivial words such as of, the, and, etc.) list.

mechanism is needed that allows searchers to "zero in" on those portions of the retrieved articles containing the hit search terms. The BRS software has two such mechanisms: (1) the Occurrence Field and (2) the "HITS" Display Option.

Occurrence Field. The occurrence field is generated during the search of the primary document file. This field records the exact location within the retrieved primary document of all search terms that caused the document to be a hit. Once generated and displayed, the occurrence field provides a means for the searcher to display those parts of the retrieved articles containing the hit search terms without having to display other parts of the articles. If the article(s) prove(s) to be of interest, any part of or the entire article may then be displayed. Without the occurrence field, the user would be faced with the time-consuming and expensive task of browsing through the entire document to find relevant information.

Figure 2 shows a typical occurrence field generated during a search. It indicates the paragraph, sentence, and word number within a retrieved document for each of the search terms that produced the hit. In the example shown in Figure 2, search terms that satisified the search logic were found in the title, abstract, five different paragraphs, and two footnotes. Once the occurrence field has been displayed, any field or combination of fields indicated in the occurrence field can easily be retrieved. The following online sessions will illustrate the procedure. The search statement requires the word THERMODYNAMIC to be adjacent to the truncated word stem PROPERT (which would retrieve either property or properties):

```
3-:    (thermodynamic adj propert$) with argon
RESULT          3
```

The search statement also requires this word combination to be in the same sentence as the word ARGON. Three primary documents were retrieved.

The following command will cause the source, title, and occurrence fields of the first document retrieved by the above search to be displayed:

```
4-:    ..browse 3 so,ti,oc/doc=1
```

Chart I

```
     1
SO ACCOUNTS OF CHEMICAL RESEARCH,VOL. 013, NO. 8, 1980, P 290- 296.
TI A VAN DER WAALS PICTURE OF THE ISOTROPIC-NEMATIC LIQUID CRYSTAL PHASE
   TRANSITION.
OC PARAGRAPH     SENTENCE  NS-WORD
   TX  (9)          3        13
   TX  (9)          3        33

            END OF DOCUMENT
```

Chart II

```
     1
TX PARAGRAPH 9 OF 36.  THE VAN DER WAALS PICTURE.  FOR EXAMPLE, WHEN PHS
   IS TAKEN FROM THE COMPUTER STUDIES OF ALDER AND WAINWRIGHT,17

   CALCULATIONS OF SEVERAL DIMENSIONLESS THERMODYNAMIC PROPERTIES, E.G.
   THE RATIO OF LIQUID TO SOLID VOLUME OR THE MOLECULAR ENTROPY OF
   FUSION IN UNITS OF BOLTZMANN'S CONSTANT, HAVE BEEN SHOWN16B TO AGREE
   VERY CLOSELY WITH EXPERIMENTAL DATA ON ARGON NEAR ITS TRIPLE POINT.
   FURTHERMORE, EQ 3 HAS BEEN DERIVED FROM "FIRST PRINCIPLES" BY KAC,
   UHLENBECK, AND HEMMER,18 WHO STUDIED THE STATISTICAL MECHANICS OF
   SYSTEMS INTERACTING VIA PAIR POTENTIALS WHICH CAN BE DECOMPOSED INTO
   A HARD-CORE REPULSION PLUS AN ATTRACTION HAVING MAGNITUDE
   +APPRX+-GAMMA+ AND RANGE +APPRX+1/+GAMMA+.  THEY SHOWED THAT EQ 3 IS
   EXACT IN THE LIMIT +GAMMA+ +FWDARW+ 0.
```

Chart III

```
     1
SO ACCOUNTS OF CHEMICAL RESEARCH,VOL. 013, NO. 8, 1980, P 290- 296.
TI A VAN DER WAALS PICTURE OF THE ISOTROPIC-NEMATIC LIQUID CRYSTAL PHASE
   TRANSITION.
AU (1) GELBART, W. M. (2) BARBOY, B.
AF (1,2) DEPARTMENT OF CHEMISTRY, UNIVERSITY OF CALIFORNIA, LOS ANGELES,
   LOS ANGELES, CALIFORNIA 90024.
TX PARAGRAPH 9 OF 36.  THE VAN DER WAALS PICTURE.  FOR EXAMPLE, WHEN PHS
   IS TAKEN FROM THE COMPUTER STUDIES OF ALDER AND WAINWRIGHT,17
   CALCULATIONS OF SEVERAL DIMENSIONLESS THERMODYNAMIC PROPERTIES, E.G.
   THE RATIO OF LIQUID TO SOLID VOLUME OR THE MOLECULAR ENTROPY OF
   FUSION IN UNITS OF BOLTZMANN'S CONSTANT, HAVE BEEN SHOWN16B TO AGREE
   VERY CLOSELY WITH EXPERIMENTAL DATA ON ARGON NEAR ITS TRIPLE POINT.
   FURTHERMORE, EQ 3 HAS BEEN DERIVED FROM "FIRST PRINCIPLES" BY KAC,
   UHLENBECK, AND HEMMER,18 WHO STUDIED THE STATISTICAL MECHANICS OF
   SYSTEMS INTERACTING VIA PAIR POTENTIALS WHICH CAN BE DECOMPOSED INTO
   A HARD-CORE REPULSION PLUS AN ATTRACTION HAVING MAGNITUDE
   +APPRX+-GAMMA+ AND RANGE +APPRX+1/+GAMMA+.  THEY SHOWED THAT EQ 3 IS
   EXACT IN THE LIMIT +GAMMA+ +FWDARW+ 0.


     2
SO THE JOURNAL OF PHYSICAL CHEMISTRY,VOL. 086, NO. 9, 1982, P
   1722-1729.
TI THERMODYNAMIC PROPERTIES OF LIQUID MIXTURES OF ARGON + KRYPTON.
AU (1) BARREIROS, S. F. (2) CALADO, J. C. (3) CLANCY, P. (4) PONTE, M.
   N. (5) STREETT, W. B.
AF (1,2,3,4,5) CENTRO DE QUIMICA ESTRUTURAL, COMPLEXO I, 1096 LISBOA,
   PORTUGAL.
```

This command requests that the source (SO), title (TI), and occurrence (OC) fields for the first document (doc=1) of answer set number 3 be displayed. The system response to this command is shown in Chart I. Notice in the occurrence field that all of the search terms that caused this document to be a hit were in the 9th paragraph of the document indicated by TX(9). To display the 9th paragraph of this document, the following command would be keyed:

```
..browse 3 tx(9)/doc=1
```

The 9th paragraph[6] of this document would then be displayed as shown in Chart II. The search terms that caused the document to be a hit are underlined here for illustrative purposes but not in the course of the actual search.

"HITS" Display Option. The BRS software also provides the "HITS" display option. This is an alternative to the above-illustrated procedure for locating portions of retrieved articles that contain search terms that caused documents to be hits. The HITS option, while utilizing the information contained in the occurrence field, does so in a manner transparent to the searcher. It allows the user to display the specific fields within the primary journal articles containing the hit search terms, but saves the searcher the steps of displaying the occurrence field and keying the command(s) necessary to display the fields containing hit search terms.

The use of the HITS display option is illustrated below for the same search in which the use of the occurrence field is illustrated above. The display command shown tells the system to display the (SO) source, (TI) title, (AU) author, and (AF)

Chart IV

```
     1
SO JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY,VOL. 028, NO. 1. 1980, P
   22- 25.
TI A COMPOSITIONAL STUDY OF PODS OF TWO VARIETIES OF MESQUITE (PROSOPIS
   GLANDULOSA, P. VELUTINA).
OC PARAGRAPH     SENTENCE  NS-WORD
    TX  (6)          3       12
    TX  (6)          3       28

     2
SO JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY,VOL. 028, NO. 5. 1980, P
   960- 963.
TI MYCOTOXIN PRODUCTION BY ALTERNARIA SPECIES GROWN ON APPLES, TOMATOES,
   AND BLUEBERRIES.
OC PARAGRAPH     SENTENCE  NS-WORD
    TX  (4)          2        5
    TX  (4)          2       12
    TX  (4)          2       33
    TX (18)          3        1
    TX (18)          3        5
    TX (18)          3       14
    TX (18)          3       15
    TX (18)          5        4
    TX (18)          5        8
    TX (18)          5       10

     3
SO BIOCHEMISTRY,VOL. 021, NO. 9, 1982, P 2036-2048.
TI ANALYSIS OF AN ALLOSTERIC BINDING SITE: THE NUCLEOSIDE INHIBITOR SITE
   OF PHOSPHORYLASE A.
OC PARAGRAPH     SENTENCE  NS-WORD
    TX  (7)          5        3
    TX  (7)          5       10
```

Chart V

```
     2
TX PARAGRAPH 4 OF 23.  THE ALTERNARIA METABOLITES WITH DEMONSTRATED
   MAMMALIAN TOXICITY BELONG TO THREE CLASSES OF COMPOUNDS: TENUAZONIC
   ACID (TEA), A TETRAMIC ACID; DIBENZO-+ALPHA+-PYRONES INCLUDING
   ALTERNARIOL (AOH), ALTERNARIOL MONOMETHYL ETHER (AME), AND ALTENUENE
   (ALT); ALTERTOXINS I AND II (ATX-I AND -II), WHICH ARE TOXIC
   SUBSTANCES OF UNKNOWN STRUCTURE.  ATX-II IS THOUGHT TO BE A DEHYDRO
   FORM OF ATX-I.  TOXICITY TO MICE FROM SPECIFIC AMOUNTS OF THESE
   COMPOUNDS HAS BEEN ESTABLISHED (PERO ET AL.  1973).  ONYALAI, A
   COMMON HEMATOLOGIC DISORDER AMONG AFRICAN BLACKS, IS CAUSED BY SALTS
   OF TEA FROM MOLD CONTAMINATION OF GRAIN (STEYN AND RABIE, 1976).
```

affiliation fields plus all the fields containing hit search terms for documents 1 and 2 (doc=1-2) of answer set number 3:

```
4-:     ..browse 3 so,ti,au,af,hits/doc=1-2
```

The display, resulting from the above command, is shown in Chart III. Hit search terms have been underlined for illustrative purposes.

ONLINE BROWSING

Once an article of interest has been found, BRS software provides the capability to electronically browse through different parts of a retrieved document. The search session will illustrate how easily a user can browse within a retrieved document:

```
1-:    (caffeine or coffee or tea) with (toxic$ or damage or adverse)
RESULT       3
```

In this search, information on toxicity of caffeine is sought. The search strategy shown requires one of the three terms in the left parentheses to be in the same sentence as one of the three terms from the right parentheses. Three documents were retrieved in this search. The following display command requests that the title, source, and occurrence fields be displayed for all three documents retrieved in the search:

```
2-:     ..browse 1 ti,so,oc/doc=1-3
```

The display generated by this command is shown in Chart IV.

The occurrence field for the second of the three documents retrieved indicates that hit search terms are in paragraphs 4 and 18 of that document. Paragraph 4 (Chart V) of this document can be displayed by using this command:

```
3-:     ..browse 1 tx(4)/doc=2
```

Having once displayed some part of a document, any other field or combination of fields can be displayed by using an abbreviated form of the browse command. For example, keying * TX(18) would generate the material shown in Chart VI.

Chart VI

```
PARAGRAPH 18 OF 23.  RESULTS AND DISCUSSION.  TEA IS USUALLY
CONSIDERED THE MOST IMPORTANT TOXIC MATERIAL PRODUCED BY ALTERNARIA
BASED ON THE POSITIVE CORRELATION BETWEEN IN VITRO PRODUCTION OF TEA
AND THE TOXICITY OF CULTURES TO LABORATORY ANIMALS (MERONUCK ET AL.
1972).  ANOTHER STUDY OF THE RELATIVE TOXICITIES OF THE ALTERNARIA
TOXINS INDICATED THAT TEA WAS SOMEWHAT MORE TOXIC THAN THE OTHER
COMPOUNDS (PERO ET AL.  1973).  THE IMPORTANCE OF TEA PRODUCED UNDER
NATURAL CONDITIONS HAS NEVER BEEN ESTABLISHED, SINCE THE ONLY
REPORTED NATURAL OCCURRENCES OF TEA HAVE BEEN IN RICE PLANT LEAVES
AND TOBACCO AND TRACE AMOUNTS IN TOMATO PASTE (SCOTT AND KANHERE,
1979).
```

Chart VII

```
     2
AU (1) STINSON, E. E. (2) BILLS, D. D. (3) OSMAN, S. F. (4) SICILIANO,
   J. (5) CEPONIS, M. J. (6) HEISLER, E. G.
AF (1,2,3,4,5,6) EASTERN REGIONAL RESEARCH CENTER, AGRICULTURAL
   RESEARCH, SCIENCE AND EDUCATION ADMINISTRATION, U.S. DEPARTMENT OF
   AGRICULTURE, PHILADELPHIA, PENNSYLVANIA 19118 (C) PRESENT ADDRESS:
   HORTICULTURAL CROPS RESEARCH LABORATORY, AGRICULTURAL RESEARCH,
   SCIENCE AND EDUCATION ADMINISTRATION, U.S. DEPARTMENT OF AGRICULTURE,
   RUTGERS UNIVERSITY, NEW BRUNSWICK, NEW JERSEY 08903.
AB KNOWN TOXIGENIC STRAINS OF ALTERNARIA ALTERNATA, ALTERNARIA
   TENUISSIMA,AND ALTERNARIA SOLANI, ISOLATED FROM A VARIETY OF PLANT
   MATERIALS, WERE CULTURED ON APPLE AND TOMATO SLICES.  WILD STRAINS OF
   ALTERNARIA SP.  WERE ISOLATED FROM COMMERCIAL BLUEBERRIES AND....

                {Remainder of abstract not shown}
```

Table IV. System Usage Statistics[a]

user category	no. of users each category	connect time (min)	av connect time per user	av connect time per session
academic chemists	44	2427	55.2	17.6
academic librarians	13	694	53.4	13.6
industrial chemists	20	1088	54.4	20.1
industrial librarians	53	2476	46.7	16.8
government chemists	10	542	54.2	16.9
government librarians	4	98	24.6	12.3
overall	144	7324	50.9	17.0

[a] Total number of users, 144; total number of online sessions, 430.

To display author names, their affiliations, and the abstract from this same article, the following command is keyed and the three requested fields are displayed as shown in Chart VII:

```
-:     * AU,AF,AB
```

RESULTS OF EXPERIMENT

During a period of approximately 3 months (August–October 1982), magnetic tapes were obtained from BRS that contained a record of user–system interactions (users' permission to do this had been requested and received). Some of the data reported in this paper were generated by a program that processed these tapes. During this period, 144 individuals participated in an experiment to evaluate the online primary journal file. The participants in the experiment were both chemists and librarians from academic, industrial, and government backgrounds. Each participant was provided with 1.5 h of free connect time by ACS. After the free connect time was used up, the participants paid for the service at the rate of $100.00 per connect hour plus approximately $7.00 per connect hour for telecommunication charges. Table IV shows the usage of the experimental file during this period, broken down by category of user.

It is interesting to note that chemists were online significantly longer each session than librarians. Is this because the chemists were less experienced than librarians in online searching or because chemists were reading the material for content or for some other reason?

Several different methods of extracting data and obtaining reactions of participants are described below. This paper contains information generated via each of these methods. (1) Several sets of questions were developed that functioned as

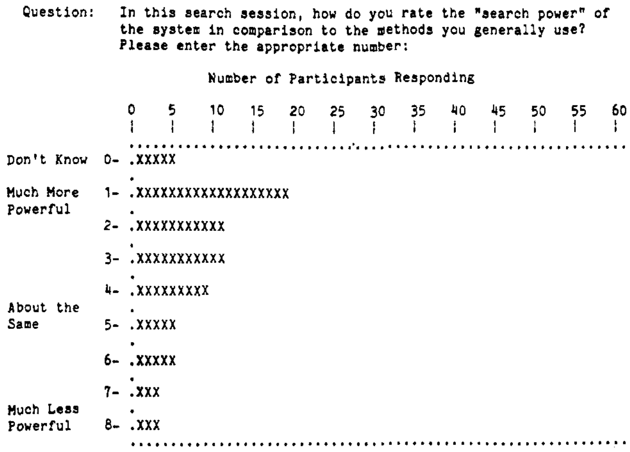

Figure 3. "Search power" of the full-text system.

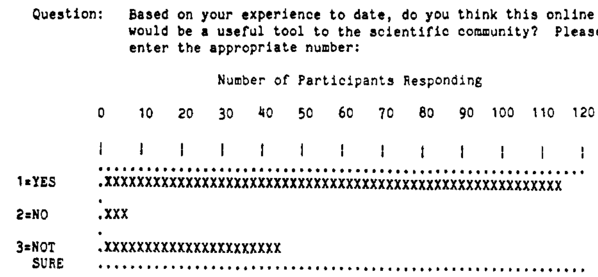

Figure 4. Utility of the full-text file to the scientific community.

an online questionnaire. Each time participants logged off, one of these sets of questions was displayed to obtain the reaction of the users while the online sessions were still fresh in their minds. (2) A computer program was written to analyze data contained on the user–system dialog tapes mentioned above. Another program was developed to tally responses to the online query sessions. (3) Following the end of the free connect time period, telephone interviews were conducted with a sampling of the participants. The participants were asked a number of questions relative to their recent experience with the full-text system. (4) Personal interview focus sessions were held with a number of the participants in the experiments. (5) Mail questionnaire surveys of participants were conducted before and after their use of the experimental file.

The sections that follow summarize the major results of the experiment. These sections are organized into various categories. The information upon which these sections are based is derived from all five data collection methods listed above.

"Search Power" of Full-Text Method. Most of the participants indicated that full-text searching was more powerful than other methods that they generally use. This searching method was thought to be particularly valuable for finding concepts embedded within the text of primary documents, which may not be the main thrust of the paper and which, therefore, might not be covered by traditional indexing methods. Figure 3 shows graphically the results of one of the questions from the online questionnaire.

The following comment relating to the search power of the full-text system was made by a participant in the experiment:

> "The full-text file is very good for searches in which the topic being sought is not the main thrust of the paper and would therefore likely be missed by indexers."

Usefulness of Full-Text File to Scientific Community. There was a strong indication from the participants that the primary journal full-text file would be an important and valuable addition to the methods of searching for chemical information. Many participants stated that the utility of the full-text file would be enhanced by the inclusion of more years of the journals available online and by extending the scope of the file to include non-ACS journals. Figure 4 illustrates graphically the responses of the participants to the question from the online questionnaire about the utility of the file to the scientific community.

The following comment relating to the utility of the system was made by a participant in the experiment:

> "It is an excellent system, but more journals are needed on the database. The system was great for finding references for writing a review article."

Economic Factors. Many participants indicated that the experimental charge of $100.00 per connect hour was too high. When the file was made available to the public in June 1983, an ACS charge (royalty) of $50.00 per connect hour was established. With telecommunication and other charges added, the total cost to the user is in the $70–90 range, depending upon the specific arrangement a given user has with BRS.

The following comment relating to the cost of the system was made by a participant in the experiment:

> "The ability to scan the entire article, including footnotes, is a strong point. The time has come for full text, but I would like to see a wider database. I would love to spend more time on the system, but the cost is too high."

Learning How To Use the System. There was a mixed response to questions about how easy or difficult it was to learn and use the query language. However, the number of participants that found the BRS language easy to learn and use was greater than the number that had difficulty. In some cases, prior knowledge of another query language seemed to be a factor for those who had problems learning the BRS language.

There was evidence that some of the participants tended to apply methods and habits from the more familiar and commonly available online bibliographic files to the full-text file. Specifically, some users retrieved titles and bibliographic data from the full-text file and then looked up the original articles in their libraries instead of browsing through the text online. This may have been due to either a desire to save connect hour charges and/or a lack of familiarity with online browsing techniques. A need for training sessions in full-text search techniques was indicated by many of the participants.

The following comment relating to the learning of full-text searching was made by a participant in the experiment:

> "I believe I used search operators that were too broad. I would have done better if I had had training sessions and a better search manual."

Online Browsing. Some of the participants liked to be able to browse online through the retrieved primary documents. The importance of being able to display and browse through the text was emphasized by some of the participants who did not have easy access to the hard-copy journals at their work locations.

The following comments relating to online browsing through the full text was made by participants in the experiment:

> "I was very much impressed with the system. I liked being able to display the paper on the screen and then print. ACS should try to expand the file beyond ACS journals and get the cost of using the file lower. I suggest subdividing full-text database by broad subject areas."

> "I used the system to search, but not to browse. I retrieved articles and then used hard copy."

Effect of Missing Graphics. As expected, absence of graphic information from the database seemed to have little, if any, effect on searchability of the full-text file. Most participants indicated that the missing graphics had only a small effect on understanding retrieved information. This is surprising and might be explained in part by the fact that the participants

Table V. Statistics on Use of Search Operators

search operator	search level	total no. of uses	uses per session
AND	document	1397	3.25
SAME	field or paragraph	437	1.02
WITH	sentence	654	1.52
ADJ	word	746	1.73
OR		538	1.25
NOT		65	0.15
$	truncation	777	1.81

did not seem to do a great deal of online browsing through the retrieved documents.

Analysis of User–System Dialog Tapes. Table V provides an indication of the usage of the various search operators. To effectively search a full-text file, it is normally necessary to use a search operator that restricts search terms to some unit of text smaller than the total article. In the BRS language, the SAME, WITH, and ADJ search operators restrict terms to the paragraph, sentence, and word level, respectively. The AND operator requires search terms to be only within the same document. While there are certain specific situations in which the AND operator must be used, most searches against a full-text file should utilize one of the more restrictive search operators to avoid large numbers of false drops.

The total usage of the three more restrictive search operators combined (SAME, WITH, and ADJ) is greater than the usage of the AND operator, as would be expected in a full-text file. However, use of the AND operator an average of 3.25 times per sessions is surprising and is probably due to unfamiliarity with full-text searching. Use of the document-level search operator is probably the result of habits formed in searching the much more common bibliographic files in which potential penalites in terms of false hits are not as great as in the case of full-text files.

CONCLUSIONS

The major conclusions of the study are summarized as follows. (1) Online full-text searching is a powerful method. The extreme depth of indexing, which is essentially total indexing, is particularly valuable in that it allows searchers to locate items of information embedded deeply within the text of primary documents that may not be the main point of the article.[7] (2) Full-text searching is a valuable addition to the information retrievel methods available to the scientific community. Utility of the file discussed here would be greatly enhanced by addition of non-ACS journals. (3) A cost of $100 per hour is too high a connect hour charge for many users, in particular users from academic institutions. (4) There was a tendency on the part of some users to apply the more familiar bibliographic searching techniques to the full-text file. To obtain good precision in full-text files, search strategies quite unlike those used in bibliographic files are usually needed. User education in full-text searching will be both useful and/or necessary.

ACKNOWLEDGMENT

We gratefully acknowledge the assistance of John T. Keys (ACS Washington) for his programming efforts in processing the data from the online questionnaire.

REFERENCES AND NOTES

(1) Schermer, C. "The Primary Journal System: A Case Study". *Graphic Commun. Comput. Assoc. J.* **1978**, *2*, 19-24.
(2) Cohen, S. M.; Schermer, C. A.; Garson, L. R. "Experimental Program for Online Access for ACS Primary Documents". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 247-252.
(3) Standard Distribution Format is described in "Chemical Abstracts Service Specifications Manual for Computer-Readable Files in Standard Distribution Format"; Chemical Abstracts Service: Columbus, OH.
(4) "BRS System Reference Manual"; Bibliographic Retrieval Services, Inc.: Latham, NY.
(5) Garson, L. R.; Cohen, S. M. "Users' Manual Primary Journal Database ACS Full-Text File"; American Chemical Society: Washington, DC, 1983.
(6) In the illustration of displayed text, there are certain words delimited by plus signs (e.g., +GAMMA+). These are spelled-out expansions of special characters that are not represented in the character set used in the online system. A list of these words is given in footnote 5.
(7) Indexing in this context refers to making all words in the text searchable, except for trivial stop words (such as of, the, and, etc.). It does not refer to the intellectual effort of indexing such as performed by abstracting and indexing services. Of course, all the ACS journals are covered by *Chemical Abstracts* as well as other services. Whether the lack of controlled indexing and standardized nomenclature within the full-text file itself is a deficiency remains to be determined.

A Relaxation Algorithm for Generic Chemical Structure Screening

ANNETTE VON SCHOLLEY†
Department of Information Studies, University of Sheffield, Sheffield S10 2TN, U.K.

Received November 7, 1983

A safe screening method for structure search within a database of generic chemical structures is described, and some results are shown. The search algorithm is based on a relaxation technique. The generic structures are represented by the Extended Connectivity Table Representation from which a data structure is set up that enables the rapid execution of the search algorithm. This data structure enables generic expressions with alternative substituents, variable positions of substituents, multipliers for singly or doubly connected substituents, and generic expressions like alkyl to be handled without need for enumeration.

INTRODUCTION

In the past a lot of quite sophisticated structure storage and retrieval systems for databases of specific chemical structures have been developed (e.g., CAS ONLINE, DARC, GREMAS). COUSIN, an online system implemented at the Upjohn Co. in Michigan, also works with a specific database, but for the query input it is possible to specify variable substituents by using the Rk notation.[1,2]

Systems dealing with generic databases, however, lag far behind those for specific structures. The Central Patent Index (CPI) of Derwent Publications,[3] Gremas of IDC (International Documentation of Chemistry),[4] and the IFI/Plenum database[5] all rely on manually assigned fragment codes.

At the University of Sheffield, a system for storage and retrieval of generic chemical structures is being developed. Generic structures are encoded in an unambiguous formal language called GENSAL,[6] which is intelligible for a chemist

† Address correspondence to the author at Beilstein-Institut, Varrentrappstr. 40-42, 6000 Frankfurt am Main 90, Federal Republic of Germany.

0095-2338/84/1624-0235$01.50/0