treated in the same manner as index terms and are assigned roles and links just as a chemical name is assigned its roles and links. Because the WLN's are also listed as broader terms, the computer will assign the appropriate WLN when the indexer encodes a chemical name. Permuting by rings has proved satisfactory and is the only degree of permuting being considered at this time.

Figure 6 summarizes the essential structure of the system.

USING THE SYSTEM

The next group of figures clarifies some of the definitions by illustrating printouts created by the system.

Figure 7 is page 142 of the thesaurus.

Figure 8 is a page from the cross-reference listing. This run picked up a duplication of notebook number 4557, and therefore N4557A was assigned to the second notebook issued as 4557 to eliminate this error. Although the system was not desinged primarily to detect errors of this type, in doing so, it gives a means for checking the filing system.

Figure 9 is page 41 of the inverted file. This lists the index terms used in the system by alphabetical order followed by the appropriate internal document numbers. If one wanted to do a manual search on malonic acid and mass spectrometers, for example, he would simply compare the internal document numbers and find that 1303 is common to both; therefore, document 1303 is a hit.

Figure 10 is a copy of the inverted-file statistics. This lists the descriptors in alphabetical order, followed by the number of documents in which this descriptor has been used, a statement about the last time the descriptor was updated, and/or deleted, and a statement of the function which the term fulfills in the thesaurus.

The Research Center System can perform four basic types of searches: document search, Boolean search, mixed search, or manual search.

The first three of these search types involve use of the computer. A manual search will normally be limited to use of the inverted file. For machine searches, there is a general list of options built into the search program. These are listed in Table I.

INVERTED FILE STATISTICS 09/15/72 PAGE 13

DESCRIPTOR	COUNT	LAST ADDITION	LAST DELETION	FUNCTION
DODD, C G	4	09/18/72		PRECISE
DODECANE	1	05/07/72		PRECISE
DODECATRICONTANE	1	09/18/72		TEMPORARY
DOTRIACONTANE	4	09/18/72		PRECISE
DOTRICONTANE-C14	8	09/18/72		PRECISE
DOVER, I C	13	09/18/72		PRECISE
DRAGENDORFF REAGENT	2	05/07/72		PRECISE
DRAWINGS	18	09/18/72		PRECISE
DRIVERS (AUTO)	1	01/23/72		PRECISE
DRYERS	1	05/06/72		PRECISE
DRYING	15	09/18/72		PRECISE
DUNN, W L, JR	1	08/28/72		PRECISE
DYTOL	22	09/18/72		PRECISE
EDMONDS, M D	43	09/18/72		PRECISE
EDWARDS, W B	16	09/18/72		PRECISE
EFFICIENCY	2	05/06/72		PRECISE
EICHORN, P A	1	01/23/72		PRECISE
ELECTRIC POTENTIAL	5	05/06/72		PRECISE
ELECTRODES	2	05/06/72		PRECISE
ELECTROLYSIS	4	08/28/72		PRECISE
ELECTROLYTIC CELLS	4	08/28/72		PRECISE
ELECTRON MICROSCOPES	5	09/18/72		PRECISE
ELECTRON PHOTOMICROGRAPHS	4	09/18/72		PRECISE
ELECTRON SPIN RESONANCE	2	09/18/72		PRECISE

Figure 10.

Table I. Search Options

(1) Specify maximum number of responses to be printed
(2) List only total number of "hits"
(3) Print, or not, index terms and text segments of document hits
(4) Truncation
(5) Specify numeric values
  a) equal
  b) not equal
  c) greater than or equal
  d) greater than
  e) less than or equal
  f) less than

LITERATURE CITED

(1) Murrill, D. P., *J. Chem. Doc.* **2**, 225-8 (1962).
(2) Murrill, D. P., *Lab. Management* **5**(6), 18-21 (1967).
(3) "EJC Thesaurus of Engineering and Scientific Terms," Engineers Joint Council, New York, N. Y., March 1969.

A Conversational Mass Spectral Search System. IV. The Evolution of a System for the Retrieval of Mass Spectral Information

STEPHEN R. HELLER,* RICHARD J. FELDMANN, HENRY M. FALES, and G. W. A. MILNE
National Institutes of Health, Public Health Service, Bethesda, Md. 20014

Received April 10, 1973

A prototype of an interactive, conversational mass spectral search system, developed at the National Institutes of Health, has been tested since September 1971 and is now being used by more than 200 scientists in the U. S. and Canada. The response has led to management of the system being given to the Mass Spectrometry Data Centre, Aldermaston, England for use by the international mass spectrometry community.

Over the past few years the Division of Computer Research and Technology (DCRT) at the National Institutes of Health has been developing prototype components of a Chemical Information System (CIS) for use by chemists and biomedical research personnel at NIH. Included in this work has been research on computer generation of Wisesser Line Notation (WLN),[1] sequential[2] and nested tree[3] structure searching, manipulation of three-dimensional structural data,[4] NMR data retrieval[5] and analysis,[6] and mass spectral data retrieval.[7-10] The last of these proj-

* To whom correspondence should be addressed at: Heuristics Laboratory, Division of Computer Research and Technology, Building 12A, Room 3001, National Institutes of Health, Bethesda, Md. 20014

ects, the mass spectral search system, has generated considerable interest and support from mass spectroscopists in the chemical and biomedical community, particularly those involved in analytical chemistry and health-related applications. As a result of this interest and through the support of the National Heart and Lung Institute, the system has been tested extensively by over 200 researchers from NIH, NIH grantees and contractors, over two dozen other government laboratories, numerous colleges, universities, medical centers, and industry.

This trial, conducted by NIH, and the Mass Spectrometry Data Center (MSDC) in England, has been quite successful and has led to an international cooperative venture between these agencies from the American and British governments. This venture provides for the transfer of the system to the GE commercial time-sharing network which will allow immediate access to local telephone numbers in 300 cities in the U. S., Canada, Europe, and Japan. All users of the system will be expected to pay for their actual computer use of the system as well as a modest yearly access fee and a royalty fee for each use of the system. These last two funds going to the MSDC will provide for updating, correcting, and expanding the data base, and possibly for running mass spectra of compounds not yet in the file. As the MSDC takes over the system, NIH will assist in an advisory capacity, when necessary, in keeping with its general policy to undertake and support research in health related areas.

SYSTEM OUTLINE AND OPTIONS

Most details of the system have been presented in detail elsewhere. At present the system options include:

1. Peak and Intensity Search
2. Molecular Weight Search
3. Complete and Partial Molecular Formula Search
4. Peak and Molecular Weight Search
5. Peak and Molecular Formula Search
6. Molecular Weight and Formula Search
7. Dissimilarity Index Comparison
8. Spectrum Printout
9. Display of Spectra
10. Plotting of Spectra
11. Microfiche Retrieval
12. CRAB—Comments and Complaints
13. HARVEST—Entering of New Data
14. NEWS—News of the System
15. MSDC Code List

The last four options, have not previously been described and are presented in this paper.

The system is readily accessed via a variety of teletypewriter compatible terminals, operating at 10, 15, 30, or 120 characters/second. To assist in the use of the system, an extensive user's manual is available.[11] All users of the system at NIH dial into Bethesda, Md. log in, and are presented with the various components of the DCRT/CIS. The start up procedure shown in Figure 1 indicates that the mass spectral search option was chosen.

Figure 2 shows the information the user provides when running the system as well as how to enter the options on the system. The user first enters his name and affiliation so that we have a record of users and their frequency of use of the system. The user then enters his three initials which are used to separate his disk scratch search files from other users who are using the search system at the same time. As many as seven users have simultaneously used the system, in addition to the regular load on the DCRT/CCB PDP-10 computer.

The CRAB option is a program which allows a user to communicate readily, quickly, and directly with those responsible for the system. Its purpose was to obtain feedback from users as effortlessly and quickly as possible, and it appears to have been successful. User comments/complaints have dealt with such items as: limitations in the size and extent of the file, troubles with dialing-up and logging on, and errors in the data base.

Of particular value to all is the last item. The greatest problems with the mass spectral data base, as with most data bases, are the errors. The interest and assistance of the users of the system to inform us of errors in the data has been most gratifying. When errors are reported and checked out, MSDC will correct them and make the corrected, searchable data base available to all users.

Figure 3 shows an example of the CRAB option and how it is used. The reason the user must restart the program (it takes about 10 seconds) is that parts of the search system are written in different computer languages—most in FORTRAN and parts in SAIL (an ALGOL dialect)—which makes automatic and proper return back to the search routine impossible.

The HARVEST option is a program to collect mass spectral data from users wishing to contribute to the sys-

```
LOG
JOB 9 DCRT/CCB 504-P TTY11
#11/226
PASSWORD:
2131   13-FEB-73   TUE

DCRT/CIS --- CHEMICAL INFORMATION SYSTEM

TO RUN A PROGRAM IN THE CIS,
TYPE THE NUMBER NEXT TO THE NAME OF THE PROGRAM

LITERATURE:
CBAC - 1

STRUCTURE:
SUBSTRUCTURE SEARCH - 2   WLN GENERATION - 3
WLN TO STRUCTURE - 4

DATA:
MASS SPEC - 5   NMR - 6

ANALYSIS:
XYZ COORDINATE GENERATOR - 7  XRAY COORDINATE REFORMAT - 8
XRAY MODELING SYSTEM - 9  XRAY CRYSTAL DATA BIBLIO SYSTEM - 10
CNDO/INDO - 11  MINDO - 12  ORTEP - 13
GINA NMR ANALYSIS - 14  ESR SPECTRUM SIMULATION - 15
MLAB CURVE FITTING/MODELING - 16

USER CHOICE: 5
```

Figure 1

```
DCRT/CIS - NHLI MASS SPECTRAL SEARCH SYSTEM

READ NEWS OF 2/5/73

PROGRAM: YOUR NAME AND COMPANY PLEASE.

USER: HELLER, DCRT, NIH

PROGRAM: PLEASE TYPE YOUR 3 INITIALS

USER: SRH

 PROGRAM: TO SEARCH FOR PEAKS, TYPE PEAK
TO SEARCH FOR MOLECULAR WEIGHT, TYPE MW
TO SEARCH FOR PEAKS WITH MF, TYPE PMF
TO SEARCH FOR PEAKS WITH MW, TYPE PMW
TO SEARCH FOR MOLECULAR FORMULA, TYPE MF
TO SEARCH FOR MW WITH MF, TYPE MWMF
TO PRINT OUT PEAKS/INTENSITIES, TYPE SPEC
TO PERFORM A DISSIMILARITY COMPARISON, TYPE SIM
TO VIEW MICROFICHE, TYPE FICHE
TO PLOT SPECTRA ON DISPLAY TERMINAL, TYPE PLOT
TO COMMENT/COMPLAIN, TYPE CRAB
TO ENTER NEW SPECTRA, TYPE DATA
TO READ THE NEWS OF THE SYSTEM, TYPE NEWS
TO LIST THE MSDC CODES, TYPE LIST
TO EXIT FORM THE PROGRAM, TYPE OUT

USER:
```

Figure 2

```
PROGRAM: PLEASE TYPE YOUR NAME AND ADDRESS AND COMMENT.
WHEN YOU ARE DONE, TYPE ZZZ AND CARRIAGE RETURN.
TO CONTINUE THE MASS SPEC SEARCH YOU MUST TYPE CHEM AGAIN.

JOHN DOE, CHEMICAL CORP. OF AMERICA, PO BOX 6472, NEW YORK, NEW YORK,10000 USA

WHEN ARE YOU GOING TO ADD LITERATURE REFERENCES TO THE SPECTRAL DATA ??

ZZZ

EXIT
```

Figure 3

tem. It is a very necessary part of the system, particularly in view of the constant CRABing regarding the size and variety of the data base. The program types out requests for the necessary information and the user responds with his spectal data. An example of the dialogue for the HARVEST option is shown in Figure 4.

```
BE NOT DECEIVED GOD IS NOT MOCKED,
    FOR WHATEVER A MAN SOWETH, THAT SHALL HE ALSO REAP.

DCRT/CIS – ALDERMASTON MASS SPEC
DATA COLLECTION PROGRAM

PROGRAM: PLEASE RESPOND TO THE QUESTIONS AS
THEY ARE PUT TO YOU. TO CONTINUE THE MASS SPEC SEARCH
YOU MUST TYPE CHEM AGAIN AFTER YOU EXIT FROM THIS PROGRAM.

MY NAME IS: JOHN DOE
MY ADDRESS (LINE 1 OF 3) IS: BLDG. 1543, CHEM. CORP. OF NEW YORK
ADDRESS LINE 2 IS: 7643 MAIN STREET
ADDRESS LINE 3 IS; NEW YORK, NEW YORK 10000
THE NAME FOR THE COMPOUND IS: CAPROLACTAM
THE CA REGN IS: 105602
THE WLN IS: T7MVTJ
THE MOLECULAR FORMULA IS: C6.H11.N.O
THE MOLECULAR WEIGHT IS: 113
THE MSDC CODES ARE: 1030, 1270
RECORDING INSTRUMENT WAS: MS-9
INLET SYSTEM: GC
INLET TEMP (C): 120
SOURCE TEMP (C): APPROX. 150
ELEC. VOLTAGE (EV): 70
ION ACCELERATING VOLTAGE (KV): 3
PRESSURE (TORR): 10-5

NOW GIVE PEAKS AND INTENSITIES

EACH LINE SHOULD CONTAIN ONE PEAK FOLLOWED BY
A COMMA AND THEN THE INTENSITY

THE PEAKS SHOULD BE IN SEQUENCE
THE INTENSITIES SHOULD BE BETWEEN C AND 100

GIVE A CARRIAGE RETURN TO FINISH INPUTTING PEAKS.
39,12
41,18
42,27
43,8
44,7
50,1
51,2
52,3
53,5
54,6
55,82
56,67
57,7
67,11
68,6
69,4
70,2
79,2
80,2
84,52
85,70
86,3
98,1
111,2
112,4
113,100
114,7

THERE WERE 28 PEAKS IN YOUR SPECTRUM

PROGRAM: DO YOU WANT TO ENTER ANOTHER SPECTRUM YES/NO

USER: NO
```

Figure 4

The NEWS option is a simple routine which prints a file stored on the disk. It contains information regarding the system, such as new options, hours of computer operations, and other system related facts that would be of value to users. It is easier and more economical to contact users this way than to use the mailing list of potential and actual users.

This list has been compiled from over 800 scientists who have responded to seminars, presentations at meetings, and notices in various publications[12-16] regarding the availability of the system. To date, over 1000 user's manuals have been distributed. The manual, in its second edition,[11] is about 60 pages, with many examples of the different options. The manual was created using the DCRT IBM 370/165 based WYLBUR text editor and is readily updated.

The last option is the MSDC Code List which, like the NEWS options, is a disk file which can be printed out. This file contains the MSDC classification codes for structural and functional group information.[17] It is useful when entering information data for the HARVEST program, so that the input will be as complete as possible.

At present, there are an average of 25 sessions per day with the system. The average cost of a search is $1 to $2, and this is expected to be 2 to 3 times higher when transferred to the commercial time-sharing system. The cost appears to be reasonable, and because of the file structure and search procedure, the cost should not increase significantly as the file grows in size.

CONCLUSION

The system would appear to have great potential and should be of value to a wide variety of users. For example, with its spectra of drugs and drug metabolites and because of its 24 hour availability, it could become an invaluable tool for hospitals all over the world in analyzing and determining the chemicals and artifacts found in drug overdose and poison cases, particularly where time is a critical factor. An example of the use of the system to aid in the saving of the life of a 6-year old farm boy in Colorado has been described recently.[18] Less dramatically, it has been found useful in reducing the (high-priced) time needed to solve structure problems since—even though the compound itself may not be present in the file—related compounds are often discovered. Furthermore, the system has been found to be of use in a pedagogical sense since mass spectral data are easily manipulated to display prominent features.

The system is scheduled to be available on the GE worldwide time sharing network in the summer of 1973. Persons interested in obtaining information as to how to use the system should address their inquiries to Dr. A. McCormick MSDC, AWRE Aldermaston, Berks, UK.

After the mass spectral search system has been in use for a reasonable period of time, the results and experience with the system will be published.

ACKNOWLEDGMENT

The authors thank A. W. Pratt for his valuable encouragement and support in making this project successful.

LITERATURE CITED

(1) Heller, S. R., and Koniver, D. A., "Computer Generation of Wiswesser Line Notation. II. Polyfused, Perifused, and Chained Ring Systems," *J. Chem. Doc.* **12**, 55-9 (1972).
(2) Feldmann, R. J., Heller, S. R., Shapiro, K. P., and Heller, R. S., "An Application of Interactive Computing: A Chemical Information System," *Ibid.*, **12**, 41-7 (1972).
(3) Feldmann, R. J., and Heller, S. R., "An Application of Interactive Graphics—The Nested Retrieval of Chemical Structures," *Ibid.*, **12**, 48-54 (1972).
(4) Feldmann, R. J., Heller, S. R., and Bacon, C. R. T., "An In-

teractive Versatile, Three-Dimensional Display, Manipulation and Plotting System for Biomedical Research," *Ibid.*, **12**, 234–7 (1972).
(5) Heller, S. R., and Feldmann, R. J., "An Interactive NMR Chemical Shift Search System," *J. Chem. Ed.* **49**, 291 (1972).
(6) Heller, S. R., and Jacobson, A. E., "GINA—A Graphical Interactive Nuclear Magnetic Resonance Analysis Program," *Anal. Chem.* **44**, 2219–22 (1972).
(7) Heller, S. R., "Conversational Mass Spectral Retrieval System and Its Use as an Aid in Structure Determination," *Ibid.*, **44**, 1951–61 (1972).
(8) Heller, S. R., Fales, H. M., and Milne, G. W. A., "An Interactive Mass Spectral Search System," *J. Chem. Ed.* **49**, 725 (1972).
(9) Heller, S. R., Fales, H. M., and Milne, G. W. A., "A Conversational Mass Spectral Search and Retrieval System. II. Combined Search Options," *Org. Mass Spectrom.* **7**, 107–14 (1972).
(10) Heller, S. R., Fales, H. M., and Milne, G. W. A., "A Conversational Mass Spectral Search System. III., Display and Plotting of Spectra and Dissimilarity Comparisons," unpublished data.
(11) Heller, S. R., "DCRT/CIS Mass Spectral Search System User's Manual," DCRT, NIH, Bethesda, Md., November 1972.
(12) *Chem. Eng. News*, pages 16–17, Oct. 16, 1972.
(13) *Mass Spectrom. Bull.*, page iii, August–December 1972.
(14) *Ind. Res.*, page 34, December 1972.
(15) *Res./Develop.*, page 44, February 1973.
(16) *J. Chem. Doc.*, **13**, 47 (1973).
(17) Scott, W., Maxwell, D. C., and Ridley, R. G., "Classification of Mass Spectral Data Files," Proc. 20th ASTM Meeting, pages 372–374, Dallas, Tex., June 1972.

Relationship between Query and Data-Base Microstructure in General Substructure Search Systems

GEORGE W. ADAMSON, VERITY A. CLINCH, and MICHAEL F. LYNCH*
Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, England

Received May 2, 1973

The distributions of bond-centered fragments which form a simple hierarchy have been investigated for a sample of substructure search queries. They are found to conform with the general pattern of fragment incidences in a file which might be used in a substructure search system. This result has been assumed in previous work on the design of screening systems for substructure search.

In previous work carried out at Sheffield on the design of screening systems for substructure search of chemical structure files,[1] it was assumed that characteristics of search requests would roughly mirror those of the file to be searched. Thus, a search for compounds containing a relatively infrequent atom, although easy to carry out, would not be requested often, whereas searches for structures containing various combinations of carbon, oxygen, and nitrogen atoms would be required much more frequently.

This assumption has been borne out by experience with fragmentation codes,[2] and we now describe our own recent work on this topic. A sample of queries was analyzed to find the incidences of fragments used as screens in the requests, and these were then compared with the incidences of such screens in the file to which the queries were to be addressed. The sample consisted of 50 user queries, supplied by the Oxford Experimental Research Unit. These were "real" queries which might be addressed to the search system. A sample of 50 queries derived from the titles of articles covered by *Current Abstracts of Chemistry* (CAC) for May 26, 1971, was also examined. These queries, as might be expected from their rather artificial derivation, were much more specific than the real queries which were usually stated in fairly general terms. This was shown by an examination of the incidences of cyclic, acyclic, aromatic, and nonaromatic bonds for the two sets of queries. The CAC queries were detailed as having alkyl and/or aromatic substituents, whereas the Oxford queries were mainly concerned with derivatives of a certain broad class of compounds, regardless of the nature of the substituent. The file used for substructure search was a random sample of 28,963 structures taken from the Chemical Abstracts Service Registry System.

Each query was analyzed in terms of differentiated simple, augmented and bonded pairs, certain species of which are used in the screen set.[1] This series of pairs forms a simple hierarchy. Some queries requested potentially varied substitution patterns, which would be covered by Boolean OR groups in a real query put to the substructure search system.[3] In these cases, all possible fragments were listed, and a value of one was assigned to each OR group, and also to each fragment which must be present (these latter are covered by AND logic). Within each OR group, equal values were assigned to equally acceptable pairs. In some cases, there were several distinct possibilities which affected likely pair fragments. For example, in $R\overset{a}{-}CH_2\overset{b}{-}X$, where R can be alkyl or aryl and X is not hydrogen, there are four possibilities for bond (a) at the augmented pair level. These are OC—C1, 1C—C1, 1C—C2, and 1C—C3. If R is aryl, 1C—C2 must be present; it may also be present if R is alkyl. Each major possibility is assigned an equal value, and the derivatives from each are then further divided. Thus, R may be aryl, in which case 1C—C2 would have the value $^1/_2$, or it may be alkyl, and the value $^1/_8$ would be assigned to each of the four possibilities. Hence, in the OR group for bond (a), 1C—C2 would have the value $^5/_8$, with OC—C1, 1C—C1, and 1C—C3 being assigned $^1/_8$ each.

* To whom correspondence should be sent.