CONCLUSION

Computerized segmentation of title words for a KWIC Index yields large benefits. For CT, these benefits include reduction of technical staff time from 4000 to 750 hr per year, reduction of typing paper and computer reports from 30 000 sheets to 8500 sheets, and reduction of computer execution time from 3 to 1 hr per year. In addition, the bibliographic data for CT are now provided from input for other CAS products. When chemists performed the segmentation process, the segmented titles were entered into the CAS data base for use in the CT KWIC Index, while the unsegmented titles were entered for use in other products. Thus one input process has been eliminated. This eliminates 7500 hr of typing and keyboarding per year.

The CAS word recognition and segmentation system is limited to chemically significant words because of the content of the dictionaries and the nature of the patterns. The system could be applied to many other disciplines by (1) selecting the types of words to include and exclude (this is the most difficult part of the work); (2) constructing a list of classified character strings in the recognized words (this would be similar to preparing a search profile with left and right truncation); (3) constructing a list of type numbers around which segmentation points should or should not fall; (4) using the applicable pattern-filling rules; (5) extracting applicable patterns and responses; (6) evaluating the applicability of the programs that remove undesirable segmentation markers.

ACKNOWLEDGMENT

CAS is pleased to acknowledge the funding of this work by the National Science Foundation (Contract C656).

LITERATURE CITED

(1) D. L. Dayton, D. R. Heym, R. Salvador, and G. M. Vanautryve, "CA Subject Index Vocabulary Standards (NSF C521 Task M and C656 III.1), Technique Description, Chemical Name Screen Algorithm", CAS Internal Report, 1972.
(2) Chemical Abstracts Service, " Substructure Searching of Computer-Readable CAS 9CI Chemical Nomenclature Files", ISBN 8412-0204-4, American Chemical Society, Washington, D.C., 1974.
(3) D. E. Knuth, "The Art of Computer Programming, Sorting and Searching", Vol. 3 Addison-Wesley, Reading, Mass., 1973, p 411.
(4) Reference 3, p 506 ff.
(5) Chemical Abstracts Service, "Facility for Integrated Data Organization (FIDO), User Reference Manual", NTIS, PB-236-020, Springfield, Va., 1974.

An International Mass Spectral Search System (MSSS). V. A Status Report

RACHELLE S. HELLER†
Computer Science Department, University College, University of Maryland, College Park, Maryland 20742

G. W. A. MILNE‡ and RICHARD J. FELDMANN‡
National Institutes of Health, Bethesda, Maryland 20014

STEPHEN R. HELLER*
Environmental Protection Agency, MIDSD, PM-218, Washington, D.C. 20460

Received April 19, 1976

The status of MSSS is described. Problems and experiences that have been encountered in three years of commercial operation of this system are reported and discussed.

INTRODUCTION

This article has been prepared to present the experiences of the international Mass Spectral Search System (MSSS) which has been operating commercially for almost three years. The MSSS is a unique system in many ways, notably in that it is a cooperative venture between two governments (U.S. and U.K.). In addition, within the U.S. government, four agencies, EPA, NIH, FDA, and NBS, are also working together toward the same goal. This structure is not free of complications, but the problems that have arisen have, for the most part, been solved, and it is felt that international collaboration of this sort is both feasible and worthwhile.

BACKGROUND

In the period 1971–1972 both EPA and NIH began to develop computer systems for aiding in the identification of compounds from their low resolution mass spectra. Both groups began by using a data base prepared by the Mass Spectrometry Data Centre (MSDC) at Aldermaston, U.K. Following some limited attempts by both EPA and NIH to disseminate their systems to the scientific community, where there was considerable interest in the MSSS, a collaborative effort was established between them and the MSDC. Under this arrangement, EPA and NIH, later joined by FDA and NBS, are funding continued development of the MSSS, while the U.K. government takes responsibility for making the MSSS available to the international scientific community. In September 1973, the MSSS was thus made available on the GE Mark III computer network. In July 1975, for technical and economic reasons, the MSSS was transferred to the ADP-Cyphernetics International Computer network.

USAGE

At present over 125 separate organizations, involving about 175 laboratories in North America and Western Europe, are using the MSSS on a daily basis. From a start of about 10 searches per day in October 1973, the rate of use has grown to over 100 per day some two years later. In addition, about 35 reference spectra from the master file are retrieved (plotted or printed) every day. User interaction with the system developers is moderate, with an average of three "CRABS" (comments or complaints written by users onto the system disk file) per week. These CRABS consist of problems, requests

† University College.
‡ National Institutes of Health.
* Environmental Protection Agency. Author to whom correspondence should be addressed.

MSSS PRICES

(FILE OF 39,509 SPECTRA)

OPTION	PRICE
PEAK	$3.00(1-5 ENTRIES) $7.00(6-10 ENTRIES)
LOSS	$4.00(1-5 ENTRIES) $7.00(6-10 ENTRIES)
FICHE	$0.25
SIM	$0.50
MW,MF	
SPEC, PLOT	$1.00
CODE, CODE/MW, BIEMANN - OVERNIGHT, LABDET	$2.00
PEAK/CODE	
PEAK/MW,MW/MF	$3.00
PEAK/LOSS, LOSS/CODE	
LOSS/MW, CODE/MF	$4.00
PEAK/MF, LOSS/MF	$5.00
BIEMANN - DAYTIME	$6.00

Figure 1. MSSS prices.

for manuals and microfiche of the file, and, most importantly, errors found in the data base.

COSTS

The U.S. government continues to fund further development of the system, and the operational costs of disk storage and maintenance are borne entirely by the U.K. government. The bulk of these costs are now covered by user subscription fees, which are $300 per year per organization ($400 during the first year). Of course, a larger file, coupled with new options, tends to raise the yearly operational costs, but the subscription fee for use of the MSSS has remained the same for over two years, during which time the file has grown from 12 879 spectra to 39 509 spectra. In addition to this annual fee, each MSSS option is priced at a fixed cost (shown in Table I) and the user must pay for connect time to the computer and the phone call to the nearest node on the commercial computer network. The Cyphernet now has nodes in over 50 cities in North America and Western Europe, and most users can access the network with a local call, thus minimizing his telephone costs. The computer is accessible 24 hr per day, using a variety of terminals such as the standard teletype and teletype compatible machines, IBM 2741, and graphics terminals, such as the Tektronix 4000 series. The computer can operate at speeds of 10, 15, 30, 120, and 200 characters per sec, depending on the user's terminal. Rental for these terminals is, of course, additional to any computer and connect charges.

The original price of a typical PEAK search, the most commonly used option, through a file of 12 879 mass spectra on the GE network was $5–$20 depending, in part, on the different pricing structures in different countries. With the universal and fixed transaction pricing scheme adopted by ADP-Cyphernetics, the cost of the same PEAK search through a file of 39 509 spectra is $3.00. Thus, in spite of inflation and a tripling in the size of the data base, the cost of a search has been reduced considerably.

RESULTS

The MSSS has now become a stable system, and this success can be traced to a number of factors.

First, no doubt, is the hard work and dedication of the large number of collaborators working on the project.

Second is the attitude of the system designers toward the users. The user, both the novice and the one with experience, has always been considered a part of the MSSS. Feedback has been encouraged, seminars and demonstrations have been given, and scientists have been encouraged to visit and consult with the various government laboratories where development of the system is carried out. In as many cases as possible, user feedback has been incorporated in the MSSS, in the form of new options (e.g., PBM and STIRS), microfiche, new and improved manuals and so on.

Third, the low costs to users of the MSSS have contributed to its acceptance and use. Since two governments were involved in the design of the system, a profit was not required from the MSSS. When it became clear that prices were too high to be commonly accepted, the software was transferred to another network and optimized to provide the service at a price which is considered reasonable and is accepted by the user community.

Last, efforts currently underway and future plans for adding new data bases (^{13}C NMR, x-ray crystal and powder diffraction) have, hopefully, given users confidence that further computer aids are coming and that by using and supporting the current effort in mass spectrometry, they are helping themselves now and for the future.

MSSS OPTIONS

The list of current and future options are shown in Table II. Most of the options have been described in the past and reference is made to those as follows:

Options	Ref	Options	Ref
1-14, 16-23	1-6	26	7
15a	8	28	11
15b, c	9, 10	29	12
24	13		

Of the more recently added options, the MSDC Bulletin literature search is the first which does not fall into the category of a program to be used as an aid in identification of unknowns. It uses a data base containing 55 000 literature references to various aspects of mass spectrometry that have appeared in the literature since 1966. It is expected that the file will be updated semiannually or annually, depending on use by and needs of the system users.

The Chemical Abstracts Service (CAS) Registry System is not really an option in itself, but rather follows from the fact that, with the next system update (August 1976) all compound names and molecular formulas will be those used in the CAS Master Registry File and will provide consistency and quality control in this area for the first time. In addition, a file of synomyns (names which are not used in the Eighth or Ninth Collective Indexes of *Chemical Abstracts*) will be added to the MSSS. Lastly, the CAS Registry Number (REGN) will be appended to every compound in the file. In this way, about 700 spectra, for which structures could not be drawn because of poor and/or ambiguous nomenclature, have already been eliminated from the data base. The REGN is also a means by which duplicate spectra can be identified and subsequently removed from the file. In fact, about one-third of the spectra have been found to be duplicates, indicating that the storage and search costs have been higher than necessary. The Wiswisser line notation (WLN) for each compound in the file, computer-generated from the CAS Registry III connection tables by software written by Gerlenter and co-workers at SUNY—Stonybrook, under contract to EPA, will also be available.

DATA BASE

The U.S. government has obtained many of the spectra in the file directly, and about 11 000 spectra that are in this category are available via the NBS/NTIS for $500. The price to foreign customers is $625. The data base will grow to about 25–30 000 spectra by late 1976. This enlarged data base will also be made available at a similar, modest cost through the NBS/NTIS.

MASS SPECTRAL SEARCH SYSTEM
(MSSS)

CURRENT AND FUTURE OPTIONS

1. Peak and Intensity Search
2. Loss and Intensity Search
3. Molecular Weight Search
4. Code Search
5. Molecular Formula Search
   (a.) Complete
   (b.) Partial, Stripped
6. Peak and Loss Search
7. Peak and Molecular Weight Search
8. Peak and Molecular Formula Search
9. Peak and Code Search
10. Loss and Molecular Weight Search
11. Loss and Molecular Formula Search
12. Loss and Code Search
13. Molecular Weight and Code Search
14. Molecular Weight and Molecular Formula Search
15. Complete Spectrum Search
    (a.) BIEMANN
    (b.) STIRS
    (c.) PBM
16. Dissimilarity Comparison
17. Spectrum/Source Print-out
18. Spectrum/Source Display
19. Spectrum/Source Plotting
20. Spectrum/Source Microfiche
21. Crab-Comments and Complaints
22. Entering New data
    (a.) Mini-Computer Interface
    (b.) Data Collection Sheets
23. News-News of the MSSS
24. MSDC Bulletin-Literature Search
25. CAS Registry Data
26. SSS-Substructure Search of CAS Data
27. WLN
28. Molecular Formula from Isotope Pattern
29. Molecular Weight from Spectral Data

Figure 2. Mass spectral search system—current and future options.

In addition to making the raw data available to the scientific community, a file of unique, CAS-registered and quality-checked data will also be made available to the public. The original quality control work was carried out by McLafferty and co-workers, under contract to EPA, and has resulted in the elimination of thousands of duplicate spectra and the correction of thousands of errors. Since the completion of this contract, over 250 000 additional corrections have been made to generate a file of the highest available quality data base. This new data base will first be made available as the MSSS file on the ADP-Cyphernetics network. Later it will be distributed to the public, but this latter step will take some time since some of the data have to be rekeypunched owing to ownership restrictions on the existing data base. Thus the intolerable situation of partial ownership of the data base being claimed by three separate organizations will be ended and, in the near future, the entire file of unique high quality spectra will be available to the public. The EPA and NIH are also publishing a book of about 30 000 spectra, with the assistance of the Chemical Abstracts Service (CAS). CAS will computer generate the book, including structural diagrams, spectra, and indexes (MF, MW, REGN and Names) using their photocomposition system.

In summary, the data base that will be made available during the summer of 1976 is expected to contain about 30 000 high-quality spectra with no duplications. In the future, updates are expected to add about 5000 new and unique spectra to the file each year.

SUMMARY AND FUTURE PROSPECTS

With the experience and positive response gained from the MSSS, we are continuing to upgrade the system as well as expand into other data bases, mainly in the area of spectroscopy. The considerable cooperation and free exchange of data with groups throughout the world indicates that future projects of this nature can also succeed, if the proper efforts and support are available.

ACCESS TO MSSS

Readers interested in obtaining further details from ADP-Cyphernetics should contact: either 175 Jackson Plaza, Ann Arbor, Mich. 48106, or J. C. Van Markenlaan 3, Postbus 286, Rijswijk (Z.H.), The Hague, The Netherlands.

ACKNOWLEDGMENT

In connection with the MSSS, we wish to express particular appreciation to the following: K. Biemann, W. F. Budde, H. M. Fales, W. Greenstreet, D. Henneberg, T. L. Isenhour, D. Koniver, D. Maxwell, J. McGuire, A. McCormick, F. W. McLafferty, J. McSorley, A. W. Pratt, R. Ryhage, M. L. Springer, V. Vinton, S. Woodward, and M. Yaguda. One of us (R.S.H.) also wishes to acknowledge the support of this international collaboration from NATO Research Grant No. 780.

LITERATURE CITED

(1) S. R. Heller, J. M. McGuire, and W. L. Budde, "Trace Organics by GC/MS", *Environ. Sci. Technol.*, **9**, 210–213 (1975).
(2) S. R. Heller, D. A. Koniver, H. M. Fales, and G. W. A. Milne, "Conversational Mass Spectral Search System. III", *Anal. Chem.*, **46**, 947–950 (1974).
(3) S. R. Heller, R. J. Feldmann, H. M. Fales, and G. W. A. Milne, "A Conversational Mass Spectral Search System. IV", *J. Chem. Doc.*, **13**, 130–133 (1973).
(4) S. R. Heller, H. M. Fales, and G. W. A. Milne, "A Conversational Mass Spectral Search System. II", *Org. Mass Spectrom.*, **7**, 107–14 (1973).
(5) S. R. Heller, "A Conversational Mass Spectral Retrieval System and Its Use as an Aid in Structure Determination", *Anal. Chem.*, **44**, 1951–1961 (1972).
(6) R. J. Feldmann, S. R. Heller, K. P. Shapiro, and R. S. Heller, "An Application of Interactive Computing: A Chemical Information System", *J. Chem. Doc.*, **12**, 41–7 (1972).
(7) R. J. Feldmann, and S. R. Heller, "An Application of Interactive Graphics—The Nested Retrieval of Chemical Structures", *J. Chem. Doc.*, **12**, 48–54 (1972).
(8) H. S. Hertz, R. A. Hites, and K. Biemann, "Identification of Mass Spectra by Computer—Searching a File of Known Spectra", *Anal. Chem.*, **43**, 681–691 (1971).
(9) K. S. Kwok, R. Venkataraghavan, and F. W. McLafferty, "Computer Aided Interpretation of Mass Spectra. III. A Self-Training Interpretive and Retrieval System", *J. Am. Chem. Soc.*, **95**, 4185–4194 (1973).
(10) G. Pesyna, R. Venkataraghavan, H. E. Dayringer, and F. W. McLafferty, "A Probability Based Matching System Using a Large Collection of Reference Mass Spectra", *Anal. Chem.*, in press.
(11) H. M. Bell, "Computer Analysis of Isotope Clusters in Mass Spectrometry", *J. Chem. Educ.*, **51**, 548 (1974).
(12) R. G. Dromey, B. G. Buchanan, D. H. Smith, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference. XIV. A General Method for Predicting Molecular Ions in Mass Spectra", *J. Org. Chem.*, **40**, 770 (1975).
(13) H. S. Hertz, D. A. Evans, and K. Biemann, "A User-Oriented Computer-Searchable Library of Mass Spectrometric Literature References", *Org. Mass Spectrom.*, **4**, 453–460 (1971).