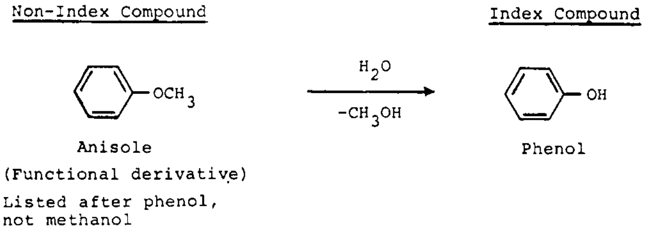


Figure 8. Index compounds and nonindex compounds in Beilstein.

(a) "CAS Today", Facts and figures about the Chemical Abstracts Service.
Chemical Abstracts Service: Columbus, Ohio, 1980.
(b) "CAS Information Tools, 1980"
Chemical Abstracts Service: Columbus, Ohio, 1980.
(c) "CAS Printed Access Tools. A workbook"
Chemical Abstracts Service: Columbus, Ohio, 1977.
(d) "This is Gmelin"
Springer-Verlag: Berlin and New York, 1979.
(e) "What is Beilstein?"
Springer-Verlag: Berlin and New York, 1978.
(f) "How to use Beilstein?"
Springer-Verlag: Berlin and New York, 1978.
(g) "Beilstein Dictionary. German-English"
Springer-Verlag: Berlin and New York, 1979.
(h) "Landolt-Börnstein Outline"
Springer-Verlag: Berlin and New York, 1978.

Figure 9. User aids.

given, since quite often this is not done by chemists themselves but by librarians and other information specialists. For those who wish to carry out interactive searches themselves, specialized user aids are required such as the "CAS Search Aid Packages". (A video tape course would have an advantage over an audio tape course in the description of an interactive search of a computer-readable file, since the dynamics of such a search can be illustrated on video tape much more readily than by means of the static illustrations of a printed audio course manual.)

The course concludes with a description of those essential peripheral publications and services that facilitate searches and information retrieval in general. Such tertiary sources and services include lists of periodicals such as the "CAS Source Index", article and tear sheet services, lists of scientists, referral centers, and buyers' guides.

Throughout the course, liberal use is made of illustrations of the search tools described. Figure 8 illustrates part of an original schematic designed to help students to see and understand the relationship between *nonindex* and *index* compounds in the Beilstein Handbook. Some of these illustrations were designed as much for experienced users of the literature and librarians as for novices, to approximate browsing through unavailable publications in order to assess probable utility.

A number of user aids are identified in the course—booklets, pamphlets, catalogs—that are distributed very often free of charge by publishers. They may be available in libraries, but may have to be specifically requested. Use of such user aids (Figure 9) may be essential since they often provide more detailed information about search tools and give more detailed directions in their use than can possibly be found in any chemical literature guide.

It is remarkable that the number of user aids has increased substantially in recent years. In the past, publishers of major secondary search tools did not appear to be visibly concerned with user problems. The trend away from this attitude is quite commendable.

A set of exercises designed to illustrate virtually all search tools described except for computer-readable ones accompanies the course. The solutions are given and in many cases these identify alternatives which underscore the partial redundancy of secondary literature sources.

An analysis of the time required by undergraduates to complete the course and to work out the exercises has been made. It is estimated that ~60 h is the total time required. This analysis stresses the fact that without access to a library, corresponding to the laboratory component in a more typical chemistry course, the course would be less effective.[8]

It remains to be seen whether these exercises will achieve the purpose of reinforcing the aims of the course and give students facility in carrying out real chemical literature searches.

REFERENCES AND NOTES

(1) Bottle, R. T., Ed. "The Use of the Chemical Literature"; Butterworths: London; (a) 2nd ed., 1969; (b) 3rd ed., 1979.
(2) Mellon, M. G. "Chemical Publications", 4th ed.; McGraw-Hill: New York, 1965.
(3) Woodburn, H. M. "Using the Chemical Literature: A Practical Guide"; Marcel Dekker: New York, 1974.
(4) Anthony, A. "A Guide to Basic Information Sources in Chemistry"; Wiley: New York, 1979.
(5) Maizell, R. E. "How to Find Chemical Information"; Wiley: New York, 1979.
(6) Arnett, E. M. "Computer-Based Chemical Information Services"; *Science* **1970**, *170*, 1370–1376.
(7) Reference 1b, p. 5.
(8) Walters, J. P.; Tylec, D. E. *Anal. Chem.* **1979**, *51*, 1233A.

NIH/EPA Chemical Information System

G. W. A. MILNE*
National Institutes of Health, Bethesda, Maryland 20502

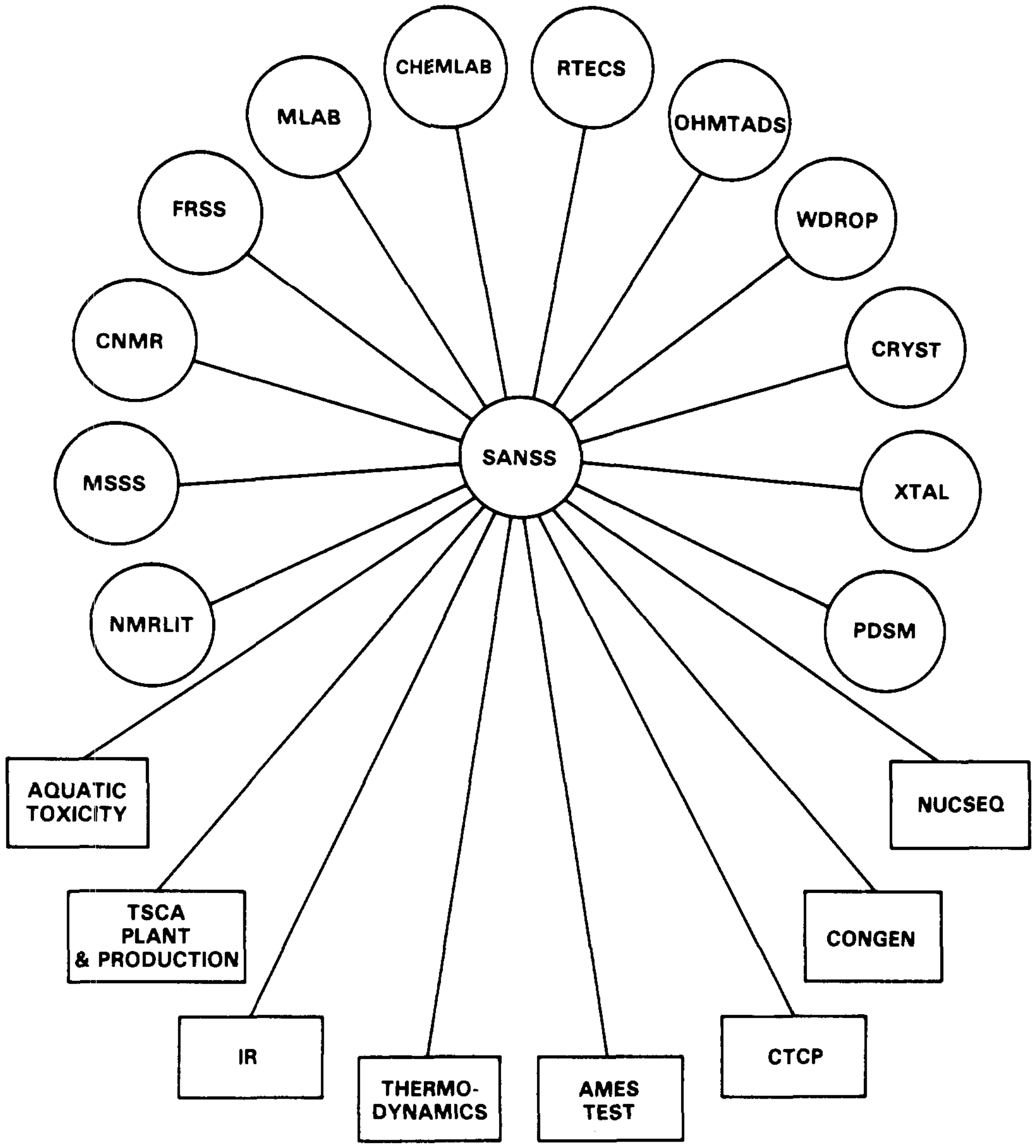
S. R. HELLER
Environmental Protection Agency, PM-218, Washington, DC 20460

Received December 13, 1979

A network of over a dozen interactively searchable chemical data bases has been made available for worldwide general use.

The NIH/EPA Chemical Information System (CIS) was started in 1973 as a joint project in mass spectrometry[1] and structure searching[2] between these two agencies and has, over the years, developed with the additional cooperation of other U.S. Government agencies,[3] as well as other organizations in the United States and elsewhere. A preliminary report of the

Current CIS Components (Spring 1980)



OPERATIONAL:	
SANSS	-Structure and Nomenclature Search System ($60/hr)
MSSS	-Mass Spectral Search System ($36/hr)
CRYST	-Xray Crystallographic Search System ($36/hr)
CNMR	-Carbon 13 NMR Search System ($36/hr)
MLAB	-Mathematical Modelling System ($60/hr)
RTECS	-Registry of Toxic Effects of Chemical Substances ($36/hr)
CAMSEQ-II	-Conformational Analysis Programs ($60/hr)
OHMTADS	-Oil and Hazardous Materials Assistance Data Base ($36/hr)
PDSM	-JCPDS Powder Diffraction Search Match ($60/hr)
FRSS	-Federal Register Search System ($60/hr)
XTAL	-Single Crystal Reduction and Search System ($36/hr)
WDROP	-Water Distribution Register of Organic Pollutants ($36/hr)
NMRLIT	-NMR Literature Search System ($36/hr)
UNDER DEVELOPMENT:	
IR	-Infrared Spectral Search System
CTCP	-Clinical Toxicity of Commercial Products
CONGEN	-Structure Generation
NUCSEQ	-Nucleic Acid Sequences

Figure 1. Schematic diagram of the NIH/EPA Chemical Information System.

prototype system was described over 3 years ago.[4]

Since that time, nine new components have been added to the Chemical Information System (CIS), regular updating schedules have been established for all CIS components, and the problems of linking between different data bases have been solved. As a result, the CIS is now best viewed as a network of chemical data bases, and it is the purpose of this paper to describe the function and utility of this system.

DESIGN OF THE CHEMICAL INFORMATION SYSTEM

A schematic representation of the CIS is shown in Figure 1. Each of the "peripheral" components consists of a data base, together with the programs for interactive searching through that data base. Identification of compounds, all of which have CAS Registry numbers, from data is possible in these peripherals.

If it is desired, on the other hand, to search for any substances with a specific structure or substructure, this can be accomplished in the central component, the Structure and Nomenclature Search System (SANSS). The entire CIS is searched by SANSS in a single operation, which retrieves the Chemical Abstracts Service (CAS) Registry number for all compounds that respond to searches. Further, SANSS informs the user as to which CIS components contain data on these compounds. The user is then free to transfer to any component where, using the CAS Registry number, the appropriate data can be retrieved and displayed.

This design has evolved in large part in response to pressures from users of the system. Its strength is that it permits one to use the system in several quite distinct ways, and it therefore is used by different people with different motives. As an example, the chemists studying water pollutants use MSSS to identify a chemical whose mass spectrum has been measured. Once the substance is identified, in terms of its CAS Registry number, information concerning its toxicity is retrieved from the National Institute of Occupational Safety and Health (NIOSH) Registry of Toxic Effects of Chemical Substances (RTECS), or spill response information is obtained from the EPA Oil and Hazardous Materials Technical Assistance Data System (OHM-TADS). These are all CIS components, from which data retrieval is a routine procedure.

Alternatively, an administrator, requiring specific information concerning a particular chemical substance, can locate that substance in SANSS, using either a systematic name or common synonym. Then the Registry number may be used to retrieve, for example, recent citations in the Federal Register for the chemical in question.

The use of unique Registry numbers to identify chemical substances leads to a simple and unambiguous system, and because of this, a decision to use the Registry number as the ultimate identifier for chemical substances within the CIS was made at an early stage of its development. A similar policy is now in effect within the EPA[5] and has been recommended as a general government policy by the Council on Environmental Quality.[6] Other systems also use this identifier, and it has thus become the link between various systems as well as between the different components of the CIS. Such a link can be exploited in a variety of ways; as an example, the Registry numbers retrieved by a substructure search in SANSS can be transferred to the Lockheed Corporation's Dialog system.[7] Here, substructure searching is based upon nomenclature and thus less precise perhaps than SANSS, but all the CAS literature citations to a compound can be easily retrieved by using its Registry number. By use of the Registry number as a link, the substructural searching power of the CIS and the enormous bibliographic scope of Dialog can be used synergistically.

NEW COMPONENTS OF THE CIS SINCE 1977

(1) Registry of Toxic Effects of Chemical Substances (RTECS). Maintained by NIOSH, this file currently contains over 70 000 toxicity measurements on some 41 000 compounds. Within the CIS, it can be searched by animal species, dosage route, and toxicity level.[8] Like all CIS files, it can be searched by chemical structure in SANSS.

(2) Oil and Hazardous Materials–Technical Assistance Data System (OHM–TADS). The Office of Water and Waste Management of EPA maintains this data base of ~1000 major chemical substances. Over 120 different types of information, ranging from toxicity data to the recommended response to spills of the substance, are available on each compound.

(3) Powder Diffraction Search–Match System (PDSM). The Joint Committee for Powder Diffraction Standards—International Data Centre has assembled a file, currently of 33 591 X-ray powder diffraction patterns, which is very useful for the identification of both organic and inorganic materials. This data base is searchable in the PDSM component of the CIS.[9] Since powder diffraction measurements are most commonly made upon mixtures of phases, PDSM uses a reverse search procedure, which checks to see if each library pattern can be "contained" within the data derived from the unknown. As a library pattern is identified within the input data, that entire pattern can be subtracted from the data and the whole search repeated upon the residue. In this way, a qualitative and semiquantitative analysis of mixtures can be accomplished.

(4) Federal Register System (FRSS). One of the critical areas of chemical information for use by regulatory agencies and commercial organizations are the chemicals cited in the U.S. Federal Register. Thus, to provide a total picture of chemical identification, chemical property (e.g., toxicity), and chemical regulations, the CIS has developed a data base and search system for the U.S. Federal Register.[10] The Federal Register Search System (FRSS) contains citations to chemicals in the Federal Register from Jan 1, 1978, to the present. The file is updated weekly and, in mid-1980, contained over 55 000 citations to chemicals found in the U.S. Federal Register. The system is searchable by chemical name, CAS Registry number, agency, type of regulation, and so forth.

(5) X-ray Single Crystal Search System (XTAL). The X-ray single crystal system is derived from the NBS/JCPDS publication titled "Crystal Data Determinative Tables". The data base will contain over 50 000 entries of crystal studies done which include information on the space group, density, and unit cell. The data base is expected to be updated annually.

(6) WaterDROP (WDROP). The WaterDROP system is a data base developed by the EPA laboratory in Athens, GA.[11] The data base is a collection of semievaluated references and citations to chemicals found in water systems in the United States and elsewhere. Hence, it serves as a Distribution Register of Organic Pollutants (DROP) found in water. At present, the system contains 10 600 citations to chemicals found in all types of water, including drinking water.

(7) Clinical Toxicology of Commercial Products (CTCP). The Clinical Toxicology of Commercial Products (CTCP), now in its fourth edition,[12] contains toxicity information for over 3000 chemicals which are part of formulations of some 20 000 products. The system, currently being tested and expected to be available for operational use in the CIS by the end of 1980, is searchable in a number of ways. The searchable parameters include chemical name(s), CAS Registry number, manufacturer, toxicity rating, and so on.

(8) Data Base Independent CIS Components. Two CIS components are processors rather than data-retrieval programs. The Mathematical Modeling Laboratory (MLAB), developed at NIH,[13] is a powerful general-purpose statistical analysis program. CHEMLAB (formerly called CAMSEQ-II[14]), which permits conformational analysis of user-defined structures, is also a CIS component.

(9) Structure and Nomenclature Search System (SANSS). Work was first started on a substructure search system (SSS) for the CIS some 10 years ago.[15] Since then the software has been continually expanded and the name of the system was changed from SSS to SANSS (Structure and Nomenclature Search System) so that the full capabilities of the system were properly recognized.[2]

Every chemical substance in the CIS has been unequivocally identified with its Chemical Abstracts Registry number much like Social Security numbers are associated, hopefully uniquely, with individuals. This process has also been completed for a considerable number of chemicals that are not in any CIS data base but which comprise integral files. Examples of this are the five substances that are regulated by Section 311 of the Clean Air Act or the 8981 chemicals which appear in the Merck Index.[16]

Each of the ~190 000 Registry numbers that has been so obtained was used to retrieve from the CAS Registry the nomenclature and structure records for the compound. These records constitute the SANSS data base.[17]

Perhaps the simplest method of searching within SANSS is with a compound name. A full name (2-chlorophenol) or a partial name (chlorophenol) may be used, the latter of course giving more hits. Any name may be truncated on the left (:orophenol, which will retrieve chloro- and fluorophenols) or on the right (chlorophen:, which will retrieve chlorophenanthrenes, for example, in addition to chlorophenols).

A higher degree of precision can usually be obtained in SANSS with a structural search. The "query structure" used in such searches must be generated by the user, with the help of the structure building commands in SANSS. Once the query structure is defined, it can be used in searches for all compounds containing a fragment centered upon a particular atom, a ring, or system of rings or the entire query structure either imbedded in a larger structure or as an exact match to the file structure.

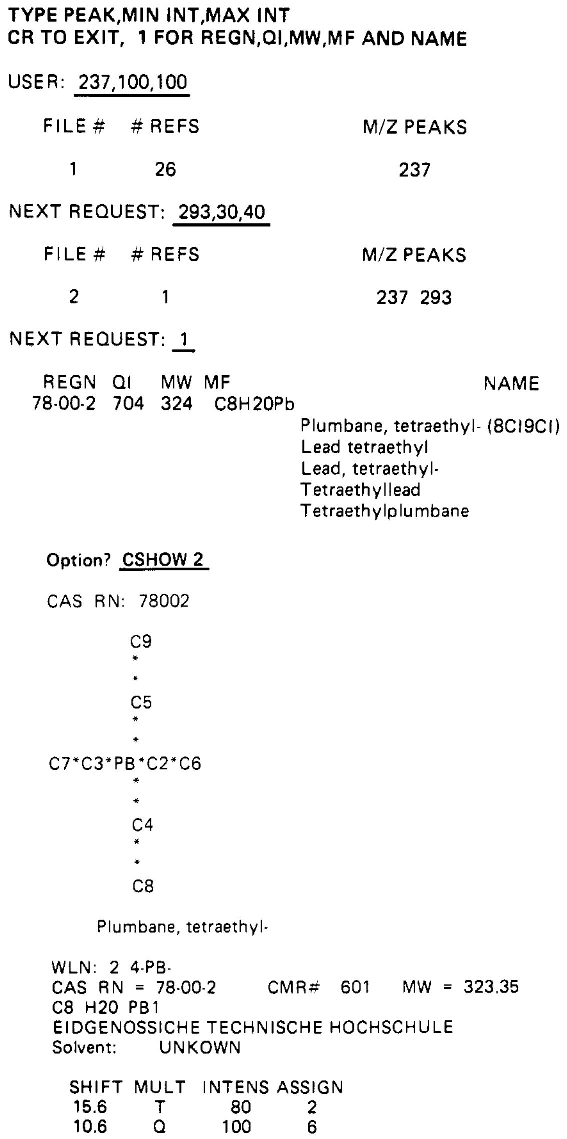
Finally, compounds containing specific functional groups can be retrieved by using specific (CIDS) keys for those groups;[18] searches for molecular weight and molecular formula are also permitted.

When a SANSS search of any sort is completed, the number of hits is reported to the user and the appropriate Registry numbers are stored in a temporary file, which can be used in subsequent Boolean operations or to display the structure and properties of the compounds that were retrieved by the search. If such a display is requested, up to five items of information concerning each compound may be printed at the terminal. These are the Registry number, the CIS data bases and non-CIS data bases that contain information pertaining to the compound and the compound's molecular formula, structural formula, and all known names and synonyms. An entire file of Registry numbers may be used by another CIS component, such as MSSS, to retrieve the mass spectra or in a non-CIS search system, such as Lockheed's Dialog,[7] to retrieve the corresponding literature citations.

EXAMPLES OF THE USE OF THE CIS

In this section, three examples are given in which the CIS was used to help solve specific problems of different sorts. The interfile linking in the system is, as will be seen, very useful in data-retrieval problems such as these.

In the first case, which is shown in Figure 2, a water pollutant has been found to exhibit in its mass spectrum intense peaks in the m/z ranges 235–237 and 293–295. The base peak (100% intensity) in the spectrum is at m/z 237, and when this is entered into the PEAK search option of MSSS,[19] a total of only 23 spectra in the data base of 32 191 are found to have their base peak at this m/z value. A second peak, at m/z 295, reduces the number of hits to two and the third peak, at m/z 293, limits the number of retrievals to just one. The pollutant is thus tentatively identified as tetraethyllead, CAS Registry No. 78-00-2. In search of confirmatory data, the carbon-13 NMR spectrum is next retrieved. This is accomplished by the command CSHOW, transfers the user from MSSS to CNMR, and then uses the Registry number to identify the correct spectrum, which is printed out.



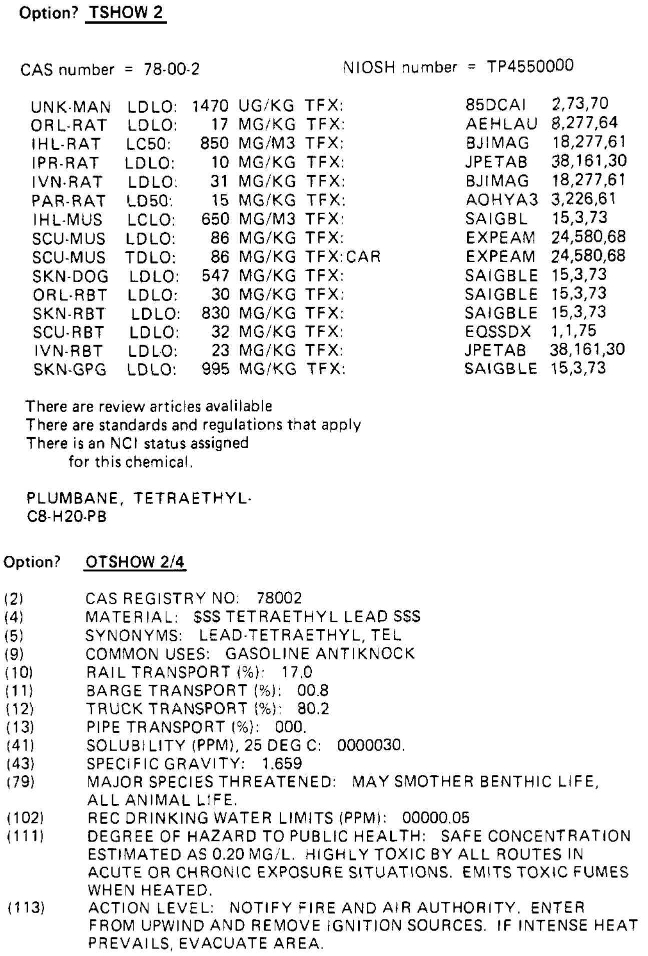
```
TYPE PEAK,MIN INT,MAX INT
CR TO EXIT,  1 FOR REGN,QI,MW,MF AND NAME

USER:  237,100,100

    FILE #   # REFS                    M/Z PEAKS

      1        26                         237

NEXT REQUEST:  293,30,40

    FILE #   # REFS                    M/Z PEAKS

      2        1                       237  293

NEXT REQUEST:  1

    REGN   QI    MW  MF                           NAME
  78-00-2  704   324   C8H20Pb
                              Plumbane, tetraethyl- (8CI9CI)
                              Lead tetraethyl
                              Lead, tetraethyl-
                              Tetraethyllead
                              Tetraethylplumbane

    Option?  CSHOW 2

    CAS  RN:  78002

              C9
              *
              *
              C5
              *
              *
    C7*C3*PB*C2*C6
              *
              *
              C4
              *
              *
              C8

          Plumbane, tetraethyl-

    WLN:  2  4-PB-
    CAS  RN  =  78-00-2      CMR#   601    MW  =  323.35
    C8  H20  PB1
    EIDGENOSSICHE TECHNISCHE HOCHSCHULE
    Solvent:      UNKOWN

      SHIFT  MULT   INTENS ASSIGN
       15.6     T        80      2
       10.6     Q       100      6
```

Figure 2. Identification by mass spectrometry and carbon-13 NMR spectroscopy of an organic pollutant.

The identification now considered fairly certain, the next task is to assess the risk implied by the discovery of this substance in a water supply. As shown in Figure 3, the command "TSHOW 1" leads to a transfer to the RTECS component. Here, lookup, retrieval, and printing of all the toxicity data available for the compound in file 1 take place, and then the user is returned to the CNMR component.

As can be seen, the substance is particularly toxic in most animal species, and so, in order to learn more about the risks posed by this substance, a transfer is requested (GO OHMTADS) to the Oil and Hazardous Materials file. Once there, the information fields pertaining to production sites (prd), degree of hazard to public health (hel), and recommended drinking water limits (drk) are sought, and the command OTSHOW 2/4 results in the retrieval and listing of all information in those fields for tetraethyllead.

From the information obtained in this session, it has become clear very quickly that this is a serious pollutant and steps to deal with this situation should be taken immediately. From initial access to the CIS to completion of the session required



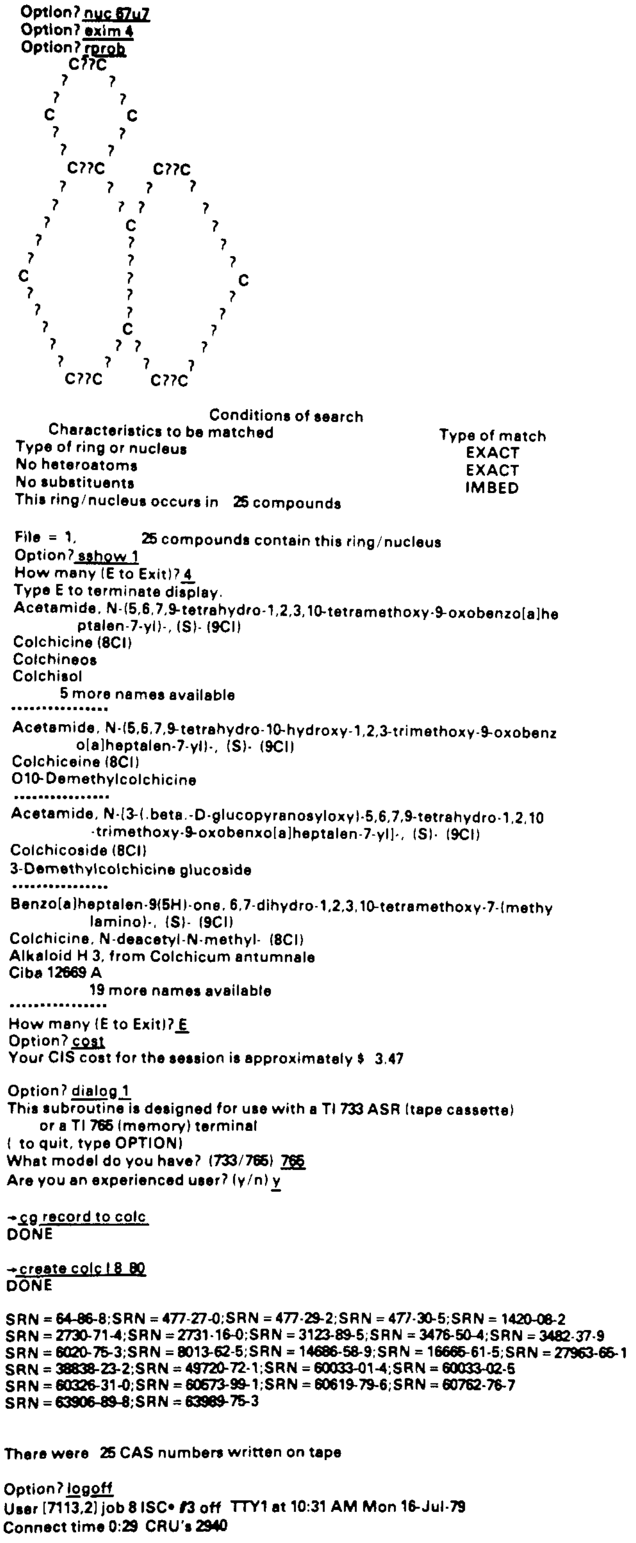
```
Option? TSHOW 2


CAS number = 78-00-2                    NIOSH number = TP4550000

 UNK-MAN  LDLO: 1470 UG/KG TFX:              85DCAI   2,73,70
 ORL-RAT  LDLO:   17 MG/KG TFX:              AEHLAU  8,277,64
 IHL-RAT  LC50:  850 MG/M3 TFX:              BJIMAG  18,277,61
 IPR-RAT  LDLO:   10 MG/KG TFX:              JPETAB  38,161,30
 IVN-RAT  LDLO:   31 MG/KG TFX:              BJIMAG  18,277,61
 PAR-RAT  LD50:   15 MG/KG TFX:              AOHYA3  3,226,61
 IHL-MUS  LCLO:  650 MG/M3 TFX:              SAIGBL  15,3,73
 SCU-MUS  LDLO:   86 MG/KG TFX:              EXPEAM  24,580,68
 SCU-MUS  TDLO:   86 MG/KG TFX:CAR           EXPEAM  24,580,68
 SKN-DOG  LDLO:  547 MG/KG TFX:              SAIGBLE 15,3,73
 ORL-RBT  LDLO:   30 MG/KG TFX:              SAIGBLE 15,3,73
 SKN-RBT  LDLO:  830 MG/KG TFX:              SAIGBLE 15,3,73
 SCU-RBT  LDLO:   32 MG/KG TFX:              EQSSDX  1,1,75
 IVN-RBT  LDLO:   23 MG/KG TFX:              JPETAB  38,161,30
 SKN-GPG  LDLO:  995 MG/KG TFX:              SAIGBLE 15,3,73

 There are review articles avalilable
 There are standards and regulations that apply
 There is an NCI status assigned
       for this chemical.

 PLUMBANE, TETRAETHYL-
 C8-H20-PB

Option?   OTSHOW 2/4

(2)       CAS REGISTRY NO:  78002
(4)       MATERIAL:  SSS TETRAETHYL LEAD SSS
(5)       SYNONYMS:  LEAD-TETRAETHYL, TEL
(9)       COMMON USES:  GASOLINE ANTIKNOCK
(10)      RAIL TRANSPORT (%):  17.0
(11)      BARGE TRANSPORT (%):  00.8
(12)      TRUCK TRANSPORT (%):  80.2
(13)      PIPE TRANSPORT (%):  000.
(41)      SOLUBILITY (PPM), 25 DEG C:  0000030.
(43)      SPECIFIC GRAVITY:  1.659
(79)      MAJOR SPECIES THREATENED:  MAY SMOTHER BENTHIC LIFE,
          ALL ANIMAL LIFE.
(102)     REC DRINKING WATER LIMITS (PPM):  00000.05
(111)     DEGREE OF HAZARD TO PUBLIC HEALTH:  SAFE CONCENTRATION
          ESTIMATED AS 0.20 MG/L.  HIGHLY TOXIC BY ALL ROUTES IN
          ACUTE OR CHRONIC EXPOSURE SITUATIONS.  EMITS TOXIC FUMES
          WHEN HEATED.
(113)     ACTION LEVEL:  NOTIFY FIRE AND AIR AUTHORITY.  ENTER
          FROM UPWIND AND REMOVE IGNITION SOURCES.  IF INTENSE HEAT
          PREVAILS, EVACUATE AREA.
```

Figure 3. Retrieval of acute toxicity and spill response information for tetraethyllead.

~4 min, and the cost of the session, as reported to the user by the COST command, was approximately $3.20.

In a quite different use of the system, it can provide a rapid means of retrieval of literature citations to specific structures or groups of structures. In the example shown in Figure 4, the problem was to locate all papers published since 1977 on compounds having the basic ring skeleton of colchicine. In SANSS, the command NUC 67u7 creates this ring system, and the option RPROBE is used to search for all compounds containing it. Any substituents are allowed and the result is 25 compounds, whose Registry numbers are stored in file 1. Four names and synonyms of a few of these are listed and can be seen to include colchicine itself, colchiceine, colchicoside, and the commercial product *N*-deacetyl-*N*-methylcolchicine. The entire file of 25 Registry numbers is next stored in the bubble memory of the Texas Instruments' Model 765 that is being used as a terminal. This is done with the command "Dialog 1", which locates the file of numbers and holds them in memory in a format appropriate for use in the Dialog system. When the storage has been completed, the user transfers through the communications network from the CIS host computer to the Lockheed Dialog host, whose network address is 41520. When a connection is made here, as is shown in Figure 5, Dialog's file no. 4 (*Chemical Abstracts*, 1977–1979) is accessed, and the Registry numbers are unloaded, five at a time, from the terminal's memory. In this way, a total of 470 citations to the 25 compounds are located, and one citation is printed out in the figure, as an example.

In a real case, of course, a sensible next step would be for the user to reduce this list of 470 citations to those of particular interest at the time. This is easily done by, for example, a subject search in Dialog, followed by the appropriate Boolean combination of files.

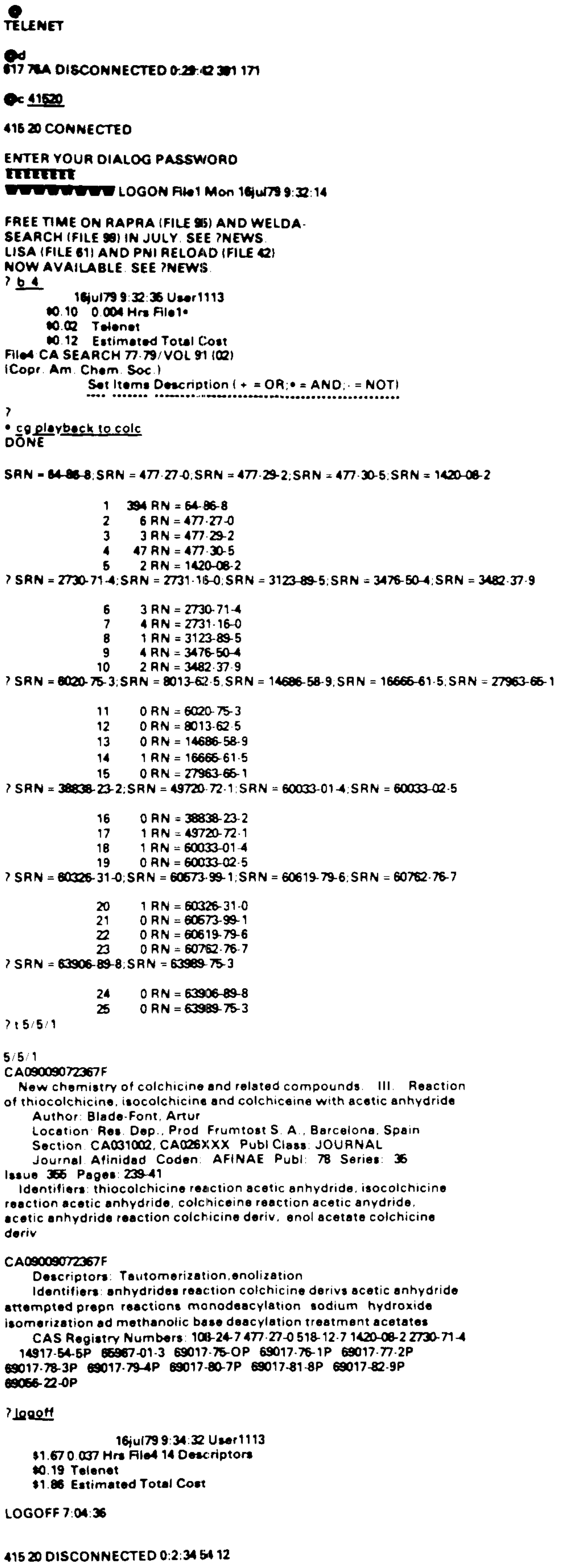


```
Option? nuc 67u7
Option? exim 4
Option? rprob
      C??C
     ?    ?
    ?      ?
   C        C
    ?      ?
     ?    ?
      C??C     C??C
     ?    ?   ?    ?
    ?      ? ?      ?
   ?        C        ?
  ?         ?         ?
 ?          ?          ?
C           ?           C
 ?          ?          ?
  ?         ?         ?
   ?        C        ?
    ?      ? ?      ?
     ?    ?   ?    ?
      C??C     C??C

                    Conditions of search
   Characteristics to be matched                    Type of match
Type of ring or nucleus                                EXACT
No heteroatoms                                         EXACT
No substituents                                        IMBED
This ring/nucleus occurs in  25 compounds

File = 1,          25 compounds contain this ring/nucleus
Option? sshow 1
How many (E to Exit)? 4
Type E to terminate display.
Acetamide, N-(5,6,7,9-tetrahydro-1,2,3,10-tetramethoxy-9-oxobenzo[a]he
        ptalen-7-yl)-, (S)- (9CI)
Colchicine (8CI)
Colchineos
Colchisol
      5 more names available
................
Acetamide, N-(5,6,7,9-tetrahydro-10-hydroxy-1,2,3-trimethoxy-9-oxobenz
        o[a]heptalen-7-yl)-, (S)- (9CI)
Colchiceine (8CI)
O10-Demethylcolchicine
................
Acetamide, N-(3-(.beta.-D-glucopyranosyloxy)-5,6,7,9-tetrahydro-1,2,10
          -trimethoxy-9-oxobenxo[a]heptalen-7-yl]-, (S)- (9CI)
Colchicoside (8CI)
3-Demethylcolchicine glucoside
................
Benzo[a]heptalen-9(5H)-one, 6,7-dihydro-1,2,3,10-tetramethoxy-7-(methy
          lamino)-, (S)- (9CI)
Colchicine, N-deacetyl-N-methyl- (8CI)
Alkaloid H 3, from Colchicum antumnale
Ciba 12669 A
          19 more names available
................
How many (E to Exit)? E
Option? cost
Your CIS cost for the session is approximately $  3.47

Option? dialog 1
This subroutine is designed for use with a TI 733 ASR (tape cassette)
    or a TI 765 (memory) terminal
( to quit, type OPTION)
What model do you have? (733/765) 765
Are you an experienced user? (y/n) y

→cg record to colc
DONE

→create colc I 8 80
DONE

SRN = 64-86-8;SRN = 477-27-0;SRN = 477-29-2;SRN = 477-30-5;SRN = 1420-08-2
SRN = 2730-71-4;SRN = 2731-16-0;SRN = 3123-89-5;SRN = 3476-50-4;SRN = 3482-37-9
SRN = 6020-75-3;SRN = 8013-62-5;SRN = 14686-58-9;SRN = 16665-61-5;SRN = 27963-65-1
SRN = 38838-23-2;SRN = 49720-72-1;SRN = 60033-01-4;SRN = 60033-02-5
SRN = 60326-31-0;SRN = 60573-99-1;SRN = 60619-79-6;SRN = 60762-76-7
SRN = 63906-89-8;SRN = 63989-75-3


There were  25 CAS numbers written on tape

Option? logoff
User [7113,2] job 8 ISC• #3 off  TTY1 at 10:31 AM Mon 16-Jul-79
Connect time 0:29  CRU's 2940
```

Figure 4. Retrieval and storage of CAS Registry numbers for all compounds with a specific ring skeleton.

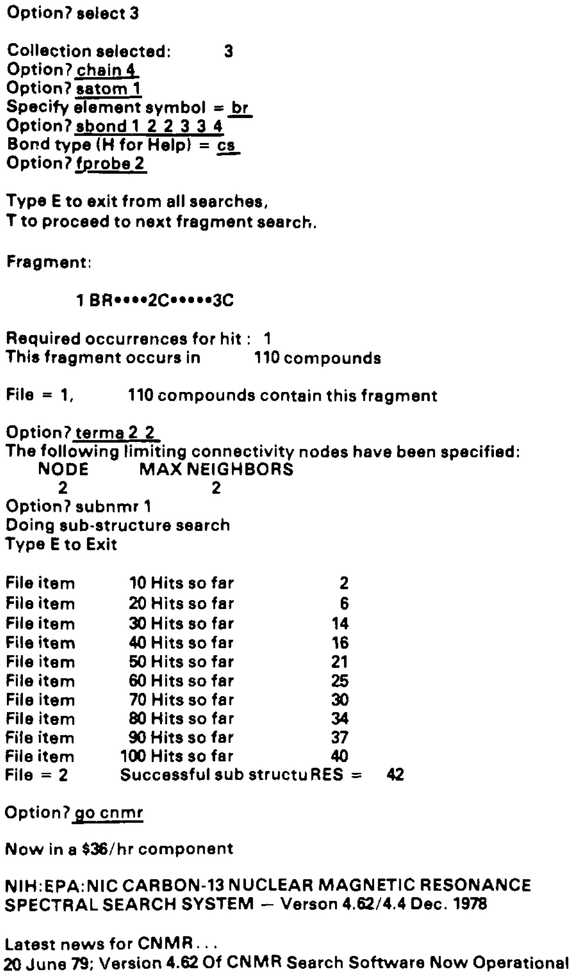
All of this searching and retrieval can be done in under 15 min, including the transfer from the CIS to Dialog. The cost within the CIS was $3.47 and the Dialog charge was $1.86 for a total cost of $4.33.

There are no other practical approaches to this sort of a problem, because the structure search in Figure 4 cannot be done manually, and attempts to conduct a search of this sort with compound names will rarely be exhaustive, as a conse-



```
@
TELENET

@d
817 78A DISCONNECTED 0:29:42 391 171

@c 41520

415 20 CONNECTED

ENTER YOUR DIALOG PASSWORD
[redacted]
[redacted] LOGON File1 Mon 16jul79 9:32:14

FREE TIME ON RAPRA (FILE 95) AND WELDA-
SEARCH (FILE 99) IN JULY. SEE ?NEWS.
LISA (FILE 61) AND PNI RELOAD (FILE 42)
NOW AVAILABLE. SEE ?NEWS.
? b 4
            16jul79 9:32:36 User1113
      $0.10   0.004 Hrs File1•
      $0.02   Telenet
      $0.12   Estimated Total Cost
File4 CA SEARCH 77-79/VOL 91 (02)
(Copr. Am. Chem. Soc.)
               Set Items Description ( + = OR;• = AND;- = NOT)
               ---- ------- -------------------------------------------

?
• cg playback to colc
DONE

SRN = 64-86-8;SRN = 477-27-0;SRN = 477-29-2;SRN = 477-30-5;SRN = 1420-08-2

                 1   394 RN = 64-86-8
                 2     6 RN = 477-27-0
                 3     3 RN = 477-29-2
                 4    47 RN = 477-30-5
                 5     2 RN = 1420-08-2
? SRN = 2730-71-4;SRN = 2731-16-0;SRN = 3123-89-5;SRN = 3476-50-4;SRN = 3482-37-9

                 6     3 RN = 2730-71-4
                 7     4 RN = 2731-16-0
                 8     1 RN = 3123-89-5
                 9     4 RN = 3476-50-4
                10     2 RN = 3482-37-9
? SRN = 6020-75-3;SRN = 8013-62-5;SRN = 14686-58-9;SRN = 16665-61-5;SRN = 27963-65-1

                11     0 RN = 6020-75-3
                12     0 RN = 8013-62-5
                13     0 RN = 14686-58-9
                14     1 RN = 16665-61-5
                15     0 RN = 27963-65-1
? SRN = 38838-23-2;SRN = 49720-72-1;SRN = 60033-01-4;SRN = 60033-02-5

                16     0 RN = 38838-23-2
                17     1 RN = 49720-72-1
                18     1 RN = 60033-01-4
                19     0 RN = 60033-02-5
? SRN = 60326-31-0;SRN = 60673-99-1;SRN = 60619-79-6;SRN = 60762-76-7

                20     1 RN = 60326-31-0
                21     0 RN = 60673-99-1
                22     0 RN = 60619-79-6
                23     0 RN = 60762-76-7
? SRN = 63906-89-8;SRN = 63989-75-3

                24     0 RN = 63906-89-8
                25     0 RN = 63989-75-3
? t 5/5/1

5/5/1
CA09009072367F
  New chemistry of colchicine and related compounds.  III.  Reaction
of thiocolchicine, isocolchicine and colchicine with acetic anhydride
    Author: Blade-Font, Artur
    Location: Res. Dep., Prod. Frumtost S. A., Barcelona, Spain
    Section: CA031002, CA026XXX  Publ Class: JOURNAL
    Journal: Afinidad  Coden:  AFINAE  Publ:  78  Series:  36
Issue  365  Pages: 239-41
  Identifiers: thiocolchicine reaction acetic anhydride, isocolchicine
reaction acetic anhydride, colchiceine reaction acetic anydride,
acetic anhydride reaction colchicine deriv,  enol acetate colchicine
deriv

CA09009072367F
    Descriptors:  Tautomerization,enolization
    Identifiers: anhydrides reaction colchicine derivs acetic anhydride
attempted prepn  reactions  monodeacylation  sodium  hydroxide
isomerization ad methanolic base deacylation treatment acetates
    CAS Registry Numbers: 108-24-7 477-27-0 518-12-7 1420-08-2 2730-71-4
  14917-54-5P  65967-01-3  69017-75-OP  69017-76-1P  69017-77-2P
69017-78-3P  69017-79-4P  69017-80-7P  69017-81-8P  69017-82-9P
69056-22-0P

? logoff

               16jul79 9:34:32 User1113
    $1.67 0.037 Hrs File4 14 Descriptors
    $0.19  Telenet
    $1.86  Estimated Total Cost

LOGOFF 7:04:36


415 20 DISCONNECTED 0:2:34 54 12
```

Figure 5. Transfer of a file of CAS Registry numbers to Dialog and retrieval of Chemical Abstracts citations.



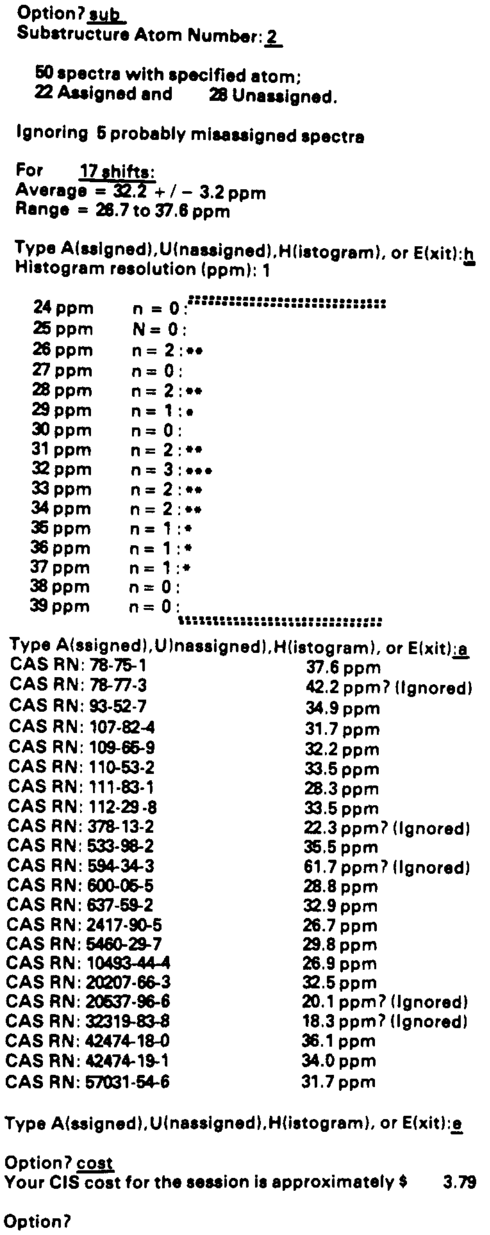
```
Option? select 3

Collection selected:        3
Option? chain 4
Option? satom 1
Specify element symbol = br
Option? sbond 1 2 2 3 3 4
Bond type (H for Help) = cs
Option? fprobe 2

Type E to exit from all searches,
T to proceed to next fragment search.

Fragment:

        1 BR••••2C•••••3C

Required occurrences for hit : 1
This fragment occurs in        110 compounds

File = 1,        110 compounds contain this fragment

Option? terma 2 2
The following limiting connectivity nodes have been specified:
    NODE        MAX NEIGHBORS
      2               2
Option? subnmr 1
Doing sub-structure search
Type E to Exit

File item      10 Hits so far          2
File item      20 Hits so far          6
File item      30 Hits so far         14
File item      40 Hits so far         16
File item      50 Hits so far         21
File item      60 Hits so far         25
File item      70 Hits so far         30
File item      80 Hits so far         34
File item      90 Hits so far         37
File item     100 Hits so far         40
File = 2       Successful sub structuRES =     42

Option? go cnmr

Now in a $36/hr component

NIH:EPA:NIC CARBON-13 NUCLEAR MAGNETIC RESONANCE
SPECTRAL SEARCH SYSTEM – Verson 4.62/4.4 Dec. 1978

Latest news for CNMR...
20 June 79: Version 4.62 Of CNMR Search Software Now Operational
```

Figure 6. Substructure search in the CNMR data base for all compounds containing the bromoethyl moiety.

quence of the inadequacies of chemical nomenclature. The cost in time and money therefore can only be compared to the alternative of a skilled scientist, browsing in the library for many hours; neither the cost of the scientist's time nor that of maintaining the library is inconsiderable, and in this sort of a comparison, the online retrieval systems, such as the CIS and Dialog, are very competitive.

A much more detailed structural search in the CIS is illustrated in a third example, shown in Figure 6, in which information is sought concerning the expected NMR chemical shift of a methylene carbon substituted by a bromine atom. The necessary substructure search is carried out in SANSS, where searching is first limited, by using the SELECT command, to the file corresponding to the CNMR data base. Then a query structure, -C-C-Br, is constructed and an FPROBE search is carried out for all compounds containing a C-Br residue. The TERMA command limits to two the number of nonhydrogen neighbors permitted for atom 2 and thus defines the substructure as -C-CH2-Br. Then the file of 110 bromo compounds is inspected with the SUBNMR search option for all compounds containing this substructure.

A total of 42 compounds are retrieved in this search, and their Registry numbers are stored in temporary file 2. The user now requests transfer to the CNMR Search System, and, once there, issues the command SUB, as shown in Figure 7. Atom 2, the carbon in the query structure bearing the bromine, is nominated, and the program uses the 42 Registry numbers in file 2, together with the corresponding CNMR data, to find that a shift for such a carbon has been assigned and reported 22 times. Five of the reported shifts are diagnosed as outliers, and from the remaining 17, the average shift is calculated as



```
Option? sub
Substructure Atom Number: 2

   50 spectra with specified atom;
   22 Assigned and      28 Unassigned.

Ignoring  5 probably misassigned spectra

For      17 shifts:
Average = 32.2 + / - 3.2 ppm
Range = 26.7 to 37.6 ppm

Type A(ssigned),U(nassigned),H(istogram), or E(xit):h
Histogram resolution (ppm): 1

   24 ppm      n = 0 :::::::::::::::::::::::::::::::
   25 ppm      N = 0 :
   26 ppm      n = 2 :**
   27 ppm      n = 0 :
   28 ppm      n = 2 :**
   29 ppm      n = 1 :*
   30 ppm      n = 0 :
   31 ppm      n = 2 :**
   32 ppm      n = 3 :***
   33 ppm      n = 2 :**
   34 ppm      n = 2 :**
   35 ppm      n = 1 :*
   36 ppm      n = 1 :*
   37 ppm      n = 1 :*
   38 ppm      n = 0 :
   39 ppm      n = 0 :
                      :::::::::::::::::::::::::::::::
Type A(ssigned),U)nassigned),H(istogram), or E(xit):a
CAS RN: 78-75-1                   37.6 ppm
CAS RN: 78-77-3                   42.2 ppm? (Ignored)
CAS RN: 93-52-7                   34.9 ppm
CAS RN: 107-82-4                  31.7 ppm
CAS RN: 109-65-9                  32.2 ppm
CAS RN: 110-53-2                  33.5 ppm
CAS RN: 111-83-1                  28.3 ppm
CAS RN: 112-29-8                  33.5 ppm
CAS RN: 378-13-2                  22.3 ppm? (Ignored)
CAS RN: 533-98-2                  35.5 ppm
CAS RN: 594-34-3                  61.7 ppm? (Ignored)
CAS RN: 600-05-5                  28.8 ppm
CAS RN: 637-59-2                  32.9 ppm
CAS RN: 2417-90-5                 26.7 ppm
CAS RN: 5460-29-7                 29.8 ppm
CAS RN: 10493-44-4                26.9 ppm
CAS RN: 20207-66-3                32.5 ppm
CAS RN: 20537-96-6                20.1 ppm? (Ignored)
CAS RN: 32319-83-8                18.3 ppm? (Ignored)
CAS RN: 42474-18-0                36.1 ppm
CAS RN: 42474-19-1                34.0 ppm
CAS RN: 57031-54-6                31.7 ppm

Type A(ssigned),U(nassigned),H(istogram), or E(xit):e

Option? cost
Your CIS cost for the session is approximately $      3.79

Option?
```

Figure 7. Retrieval and statistical analysis of all carbon-13 chemical shifts for the bromoethyl moiety.

32.2 ppm, with a standard deviation of 3.2 ppm.

The cost of this complete inquiry is $3.79, and it was completed in a little more than 5 min. Again, this is a search which is almost impossible manually; the compilation of tables of CNMR shifts, which contain items such as this value, has required many expert spectroscopists to spend considerable time, and while such tables are very valuable, they do not cater to the chemist who is interested in the likely chemical shifts for the carbons of a particular unusual structure. This SANSS/CNMR linked search can tolerate any structure; its limitation is that the CNMR data base, for reasons discussed in the next section, is not large enough to support wide-ranging queries.

MANAGEMENT ASPECTS OF THE CIS

Development and operation of the publicly available CIS have been in progress for about eight years now, and during that time, it has become increasingly clear where the major difficulties lie. This section contains a discussion of some of these points, consideration of which is relevant to continuing development of the system.

(1) Programming and Data Acquisition. In the early days of the project, the cost of software development was considerable, but as generalized software is being implemented, the proportion of total development costs that support programming is steadily diminishing. By contrast, the cost of data collection and evaluation has emerged as the largest continuing cost within CIS development and also one of the most intractable problem areas, because it involves both data acquisition and evaluation, both of which present special difficulties.

Much of the early collection of data for the CIS was done on an ad hoc basis in that mass spectroscopists, for example, were contacted and asked to provide spectra for the data base. Most of the known collections of spectra have been acquired and merged into the growing file and, consequently, this approach has in recent years become less fruitful. It also has the added disadvantage that it allows little control over the content of the file. For these reasons, mass spectra are now being acquired by direct measurement on those compounds whose presence in the data base is felt to be important. This approach permits some control of the file content and quality but is very costly; acquisition and entry of a mass spectrum into the file in this way can cost as much as $250.

The primary literature provides very little assistance in problems of this sort. This is generally because the cost of page space is now so high that most journals refrain from publication of actual data. Thus, those interested in collection of data are in the tantalizing position of knowing that data exist but are not directly available. This is a gloomy situation, relieved only by the light shed upon it by a program announced in 1977[20] by the Chemical Society in London. Those wishing to publish X-ray crystallographic papers in the journals of the Chemical Society are now required to submit all numeric data together with the manuscript. The data may be reviewed by referees but are generally not published in the journal; instead, they are sent to the Crystal Data Centre at Cambridge University where they are used in the compilation of the Cambridge Crystal file, an authoritative data base, which is leased from the Data Centre by most countries in the Western world. The income from these leases supports the Data Centre, and the cost of acquiring the raw data is reduced to almost zero. This experiment appears to have been quite successful, but it has yet to be emulated by other chemical societies.

The field of data evaluation is very poorly developed, and there is general agreement only that this is an important activity if data bases are to be reliable. The cost of data evaluation is, however, considerable, and in many cases, it is not even clear how data can be evaluated, at any cost, without actually repeating the measurements in question. There have been some tentative efforts directed at validation of data in the CIS files, and the experience of the National Bureau of Standards has proved to be very helpful in this area. It remains a fact, however, that most of the data in the CIS is unevaluated, and how it might be evaluated economically is not obvious.

(2) User Support and Marketing. The CIS currently has more than 900 users from over 400 user organizations who carry out between 30 000 and 40 000 transactions per month. Consequently, it is important that there be adequate communication between them and the system managers, and a number of approaches have been adopted to ensure that this is the case.[21]

A monthly CIS newsletter is published and sent to over 2000 individuals. This publication, which is typically about four to six pages in length, provides general information about the CIS. Those who use the system have access to an online news system, which provides headlines at login and the full news message upon request. This news system is the best means of ensuring the users are apprised in detail of changes in the CIS.

As far as communication from users and others is concerned, two procedures have proved effective. A toll-free telephone is maintained on a 24-h basis and is used perhaps 12 times per day, mostly by nonusers of the CIS, for inquiries of various sorts. Users of the system can submit comments or complaints using a CIS option, called COMMENT. Such comments are collected and reviewed daily. On average, about one comment per day is received and the appropriate response is provided as soon as is possible, typically within 3 days.

(3) Future Directions. One of the major administrative concerns with the CIS has been its cost and implicit open-ended nature. The CIS was designed to be a public system, and it was decided that an annual $300 subscription fee for access to the system should be levied. (This fee has been waived for educational institutions as of May 1980.) Use of a commercial computer is consistent with OMB Circular A76[22] and leads to the generation of funds which is used to defray the cost to the government of the user support discussed in the preceding section. Current projections suggest that the subscription fee and other revenues will match the support and marketing costs in 1981. After this time, these revenues will be used to defray the costs of maintenance of the CIS. In this way, provided use of the CIS does not decrease (which is hardly likely given the recent doubling of growth per year for the last two years), it is possible that it could become self-supporting at the operational level by 1982.

As these efforts to control costs are continuing, some attempts are being made to control the directions in which the CIS is growing. A Steering Committee now reviews questions such as addition of new components and modification of existing modules, and, as this Committee is open to all the many groups, governmental and otherwise, which have collaborated in the development of the system, it is hoped that it will function democratically in guiding the future development of the CIS. Lastly, a Science Advisory Board, comprised of scientists from industry and academia, has been established to assure that the highest possible scientific levels are maintained in the system.

Thus, the CIS has evolved to the point where a solid system has been built and is being used. The tasks ahead now are to develop a smoothly functional maintenance and operation staff so that the main government efforts can be directed to further developments in areas deemed relevant to those senior-level decision-making managers who control the funding of such projects as the CIS.

REFERENCES AND NOTES

(1) S. R. Heller, *Anal. Chem.*, **44**, 1951 (1972); S. R. Heller, H. M. Fales, and G. W. A. Milne, *Org. Mass Spectrom.*, **7**, 107 (1973); S. R. Heller, H. M. Fales, and G. W. A. Milne, *J. Chem. Educ.*, **49**, 725 (1973); S. R. Heller D. A. Koniver, H. M. Fales, and G. W. A. Milne, *Anal. Chem.*, **46**, 947 (1974); S. R. Heller, R. J. Feldmann, H. M. Fales, and G. W. A. Milne, *J. Chem. Doc.*, **13**, 130 (1973); S. R. Heller, H. M. Fales, G. W. A. Milne, R. J. Feldmann, N. R. Daly, D. C. Maxwell, and A. McCormick, *Adv. Mass Spectrom.*, **6**, 1037 (1975); S. R. Heller, G. W. A Milne, R. J. Feldmann, and R. S. Heller, *J. Chem. Inf. Comput. Sci.*, **16**, 176 (1976); G. W. A. Milne and S. R. Heller, *Am. Lab.* (*Fairfield, Conn.*), **8**, 43 (1976); S. R. Heller, R. S. Heller, A. McCormick, D. C. Maxwell, and G. W. A. Milne, *Adv. Mass Spectrom.*, **7B**, 985 (1977).

(2) R. J. Feldmann, G. W. A. Milne, S. R. Heller, A. Fein, J. A. Miller, and B. Koch, *J. Chem. Inf. Comput. Sci.*, **17**, 157 (1977).

(3) These Agencies were the National Institutes of Health, the Environmental Protection Agency, the National Bureau of Standards, the National Institute of Occupational Safety and Health, and the Food and Drug Administration. In addition support from the U.K. Department of Industry has been made available since 1971.

(4) S. R. Heller, G. W. A. Milne, and R. J. Feldmann, *Science*, 195, 253 (1977).

(5) EPA Internal Order 2800.2 states that "Any computer-based Agency ... system ... containing data ... on ... chemical substances shall contain the CAS Registry number for each chemical substance...." See ref 6, p 38.

(6) "The Feasibility of a Standard Chemical Classification System and a Standard Chemical Substances Information System", report to the Congress from the Council on Environmental Quality, Stock No. 041-011-00039-4, U.S. Government Printing Office, Washington, D.C., July 1978.

(7) The Lockheed Corporation serves as a vendor for the CAS Condensate and CASIA data bases. Dialog, a Trademark registered with the U.S. Patent and Tradmark Office, is a proprietary product of Lockheed Information Systems Inc., Palo Alto, CA 94304.

(8) J. R. McGill, S. R. Heller, and G. W. A. Milne, *J. Env. Path.* and *Tox.*, **2**, 539, (1978).

(9) R. G. Marquart, I. Katsnelson, G. W. A. Milne, S. R. Heller, G. G. Johnson, and R. Jenkins, *J. Appl. Crystallogr.*, **12**, 629 (1979).

(10) G. Marquart, L. Marquart, S. Mintz, J. McGill, J. McDaniel, S. R. Heller, and G. W. A. Milne, *Online* (*Weston, Conn.*), **4**, 45 (1980).

(11) A. Alford, R. Potenzone, S. R. Heller, S. Mintz, and G. W. A. Milne, publication in preparation.

(12) R. E. Gosselin, H. C. Hodge, R. P. Smith, and M. N. Gleason, "Clinical Toxicology of Commercial Products", 4th ed., Williams and Wilkins, Baltimore, MD, 1976.

(13) G. D. Knott and R. I. Shrager, *Assoc. Comput. Mach.*, SIGGRAPH Not. **6**, 138 (1972).

(14) R. Potenzone, Ph. D. Thesis, Case Western Reserve University, February 1979. See also R. Potenzone and A. J. Hopfinger, "Structural Correlates of Carcinogenesis and Mutagenesis. A Guide to Testing Priorities", Proceedings of the 2nd FDA Office of Science Summer Symposium, Aug 28, 1978.

(15) R. J. Feldmann and S. R. Heller, *J. Chem. Doc.*, **12**, 48 (1972).

(16) "The Merck Index", 9th ed. Merck & Co., Inc., Rahway, NJ 07065, 1977.

(17) G. W. A. Milne, S. R. Heller, A. E. Fein, E. F. Frees, R. G. Marquart, J. A. McGill, and J. A. Miller, *Comput. Sci.*, **18**, 181 (1978).

(18) "Handbook of CIDS Chemical Search Keys", Fein-Marquart Associates, Inc., Baltimore, MD 21212, Nov 1973.

(19) S. R. Heller, *Anal. Chem.*, **44**, 1951 (1972).

(20) See the Notice to Authors, *J. Chem. Soc., Chem. Commun.*, **No. 3** (1977), **No. 3** (1978).

(21) For details on the CIS operations and availability of the CIS, please contact Kay Pool, CIS Project, Information Science Corporation, 2135 Wisconsin Avenue, NW, Washington, DC 20007 [(202) 298-6200 or (800) 424-2722].

(22) "Circular A76" (Office of Management and Budget, Washington, DC, Aug 1967) states, in essence, that all attempts should be made to ensure that systems such as the CIS, operate in the private sector.