of a superatom (in the form of a submatrix) can be determined, and the superatoms can be properly embedded into a structure without duplicates. Also by introducing some constraints the generated structures can be pruned to a few candidates. Therefore our algorithm can serve as a practical structure generator, but has the advantage of being simple and can even be executed on many mini/microcomputers.

ACKNOWLEDGMENT

We are particularly indebted to Professor Joseph San Filippo, Jr. of Rutgers University and Dr. Dennis H. Smith of Standford University for their comments in preparing this manuscript. We acknowledge partial financial support by NSF (Grant 80-17045) and DOE (Contract DE-AS05-80ER-10662).

REFERENCES AND NOTES

(1) Rouvray, D. H. *Chem. Br.* **1974**, *10*, 11.
(2) "Chemical Applications of Graph Theory"; Balaban, A. T., Ed.; Academic Press: New York, 1976.
(3) Masinter, L. M.; Sridharan, N. S.; Carhart, R. E.; Smith, D. H. *J. Am. Chem. Soc.* **1974**, *96*, 7702.
(4) Morgan, H. L. *J. Chem. Doc.* **1965**, *5*, 107.
(5) Randic, M. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 171.
(6) Schubert, E.; Ugi, I. *J. Am. Chem. Soc.* **1978**, *100*, 37.
(7) Smith, D. H. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 203.
(8) This process can also be carried out by permutation, but the procedure is rather lengthy.
(9) A referee has pointed out that the most current (but as yet unpublished) version of the CONGEN program has incorporated a structure generator which functions in a manner very similar to that described in this paper: Carhart, R., unpublished results.
(10) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. *J. Am. Chem. Soc.* **1975**, *97*, 5755.
(11) Sasaki, Shin-Ichi; Abe, Hidetsugu; Hirota, Yuji; Ishida, Yoshiaki; Kudo, Yoshihiro; Ochiai, Shukichi; Saito, Keiji; Yamsaki, Tohru *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 211.
(12) Shelley, C. A.; Woodruff, H. B.; Snelling, C. R.; Munk, M. E. In "Computer-Assisted Structure Elucidation"; Smith, D. H., Ed.; American Chemical Society: Washington, DC, 1977; p 92.
(13) Copies of the program listing and appropriate documentation are available upon request.

The Development of an Environmental Fate Data Base

PHILIP H. HOWARD,* GLORIA W. SAGE, and A. LAMACCHIA
Syracuse Research Corporation, Syracuse, New York 13210

ANDREW COLB
U.S. Environmental Protection Agency, Washington, DC 20460

Received September 16, 1981

Three components (DATALOG, XREF, and CHEMFATE) of a new Environmental Fate Data Base are described. Environmental fate is a term describing the behavior (i.e., transport and degradation) of a chemical which is released to the environment. The system stores and retrieves data and provides sufficient flexibility to meet the needs of a variety of environmental fate data users. It is anticipated that both government and industry will find the information in these files useful as a source of data for estimations and modeling of environmental fate and exposure evaluations, as well as structure–reactivity, persistence, or transport correlations. These correlations are particularly desirable for predicting environmental behavior of chemicals for which only a limited amount of enivronmental fate data are available. The CHEMFATE data base will also be useful in determining where research effort is needed to supply missing data on physicochemical properties and environmental degradation and transport behavior.

INTRODUCTION

With the growing awareness of the health and environmental hazards associated with the commercial production, use, and disposal of industrial chemicals, risk assessment has become an area of increasing concern and activity.[1] There are two major factors that have to be considered for an overall environmental risk assessment:[2] (1) exposure and (2) toxicity. Numerous references are available containing tabulated biological effects data,[3-6] some of which are available online;[3,7-9] however, in contrast, few tabulations of data relevant to environmental exposure (e.g., environmental release and environmental fate) exist other than a limited number of monitoring data bases.[10] This is particularly true of the substantial amount of environmental fate (i.e., transport and degradation) information that is available. In November 1979, the development of an Environmental Fate Data Base was initiated in an attempt to fill this gap.

The knowledge of how a chemical will behave in the environment once it is released is particularly important in determining whether a chemical will come in contact with a critical species or with man in sufficient concentrations to cause a toxic effect or, in contrast, be rapidly degraded to innocuous products. The type of information pertinent to the fate of a chemical released into the environment is diverse and includes physical and chemical properties, transport and degradation studies, ambient monitoring data, and field studies.[11] Considerable amounts of time and money must be expended to extract this type of data from primary literature sources. Thus, once obtained, these data should be stored in a form that is readily accessible to other investigators. A data bank of environmental fate information serves the following purposes:

(1) Allows rapid access to all available fate data on a given chemical without having to resort to expensive, time-consuming, and inefficient searches of the primary literature.
(2) Identifies critical gaps in the available information to facilitate planning of research needs.
(3) Provides a source of data (training set) for constructing structure–activity correlations for degradability and transport of chemicals in the environment. Such correlation would be a tremendous aid in identifying persistent chemical classes as well as physical or chemical properties that may correlate to particular behavior in the environment.

SYSTEM OVERVIEW

The Environmental Fate Data Base is comprised of three interrelated files called DATALOG, XREF, and CHEM-

0095-2338/82/1622-0038$01.25/0

Table I. Chemical Classes Included in the Test Set of 200 Chemicals

- hydrocarbons (HC)
  - aliphatic, acyclic
  - aliphatic, cyclic
  - aromatic
- oxygen-containing HC
  - alcohols and polyols
  - phenolics
  - aldehydes and ketones
  - ethers, acyclic
  - ethers, cyclic
  - carboxylic acids and salts
    - monoacids
    - diacids
  - acid esters and anhydrides
  - peroxides
- halogen-containing organics
  - aliphatic HC halides
  - aromatic HC halides
  - halogenated alcohols and phenols
  - haloethers
  - halo acids and derivatives
  - acyl halides
- nitrogen-containing compounds
  - amines
    - acyclic monoamines
    - cyclic monoamines
    - diamines and triamines
    - substituted amines
  - amides and imides
  - ureas, uracils, and guanidines
  - nitro compounds
  - azenes, azo compounds
  - cyanides, cyanates, isocyanates, hydrazines, oximes
  - carbamates
- sulfur-containing compounds
  - sulfides, sulfones, sulfoxides
  - sulfonic acids, sulfuric acids, and derivatives
  - thiocarbamates, dithiocarbamates, and thiocarboxylic acids
  - sulfates
- phosphorus-containing compounds
- others

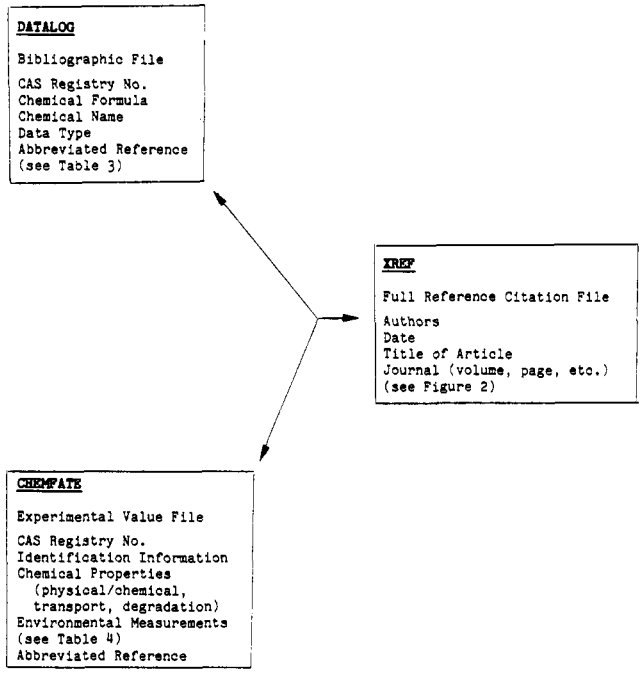

Figure 1. Organization of files for the environmental fate data base.

FATE (Figure 1). DATALOG is a data index file that indicates the types of fate data available on a particular chemical and the corresponding citations keyed by CAS registry number. Large amounts of index information on chemicals can be incorporated into the file rapidly, since individual articles require only cursory examinations for data type. No experimental values are entered into the file. In contrast to DATALOG, CHEMFATE is the data value file. Because actual experimental values (rate constants, experimental conditions, physical properties, etc.) are abstracted and retained in the file, data entry into CHEMFATE is considerably more time-consuming and exacting. Both the DATALOG and CHEMFATE files are linked to XREF, the full citation file, through an abbreviated reference [e.g., Smith, A. E. et al. (1976)]. The DATALOG file is keyed by the CAS registry number and the XREF file is keyed alphabetically by author. This structure allows rapid retrieval of information when searching by these keys but considerably slower retrieval by searching other terms. The CHEMFATE file is constructed of multiline records which always contain the CAS Registry number, data type, reference, and a record number. A parallel log is maintained which also contains these terms and allows for multiple key indexing for rapid searching by these terms. At present, these files are independently maintained and stored in the main computer at the Syracuse Research Corporation with support from the U.S. EPA Office of Pesticides and Toxic Substances. Data manipulation and report making programs have been developed at that facility.

Table II. Citation Data Bases Examined for Environmental Fate Data

Chemical Abstracts[a]
National Technical Information Service (NTIS)[a]
SciSearch
Pollution Abstracts
Ocean Abstracts
Enviroline
Air Pollution Technical Information Control (APTIC)
Environmental Bibliography
Agricola
Aquatic Science and Fisheries Abstracts

[a] Data bases which are most effective in providing environmental fate data.

DATA ACQUISITION

Because of the diverse nature of the data pertinent to the environmental fate of a chemical, it was decided that the best way of evaluating the utility of an environmental fate data base and the amount of fate data available was to collect data on a test set of 200 chemicals. The chemicals were selected from a list of 600 chemicals whose annual production exceeds one million pounds.[12] The 600 chemicals were organized by chemical classes (Table I), and then 200 chemicals were chosen in cooperation with the EPA by selecting a variety of structures in a given class. Literature searches were conducted manually and online on a number of citation data bases (Table II).

Chemicals are initially searched in the Chemical Abstracts bibliographic citation files by CAS number. In searches which generate over a hundred citations, the search is narrowed further by means of a textual search strategy which consists of a group of search terms such as biodegradation, photooxidation, etc. To assure search continuity and uniformity, this online search strategy is maintained as a "Save Search" in the Dialog System. In addition, the citation file may be further refined through combination (AND in Boolean logic) with selected Chemical Abstracts section codes (3–5, 10, 12,

Table III. Example of DATALOG Record for Rapid Storage of Retrieved References

CAS no.	formula	chem name	data type[a]	ref
100-02-7	C6H5N03	4-NITROPHENOL	ADSORP.	CALLAHAN, MA ET AL. (1979A)
100-02-7	C6H5N03	4-NITROPHENOL	BIOCON.	CALL, DJ ET AL. (1980)
100-02-7	C6H5N03	4-NITROPHENOL	BIOCON.	HOWARD, PH ET AL. (1976)
100-02-7	C6H5N03	4-NITROPHENOL	BIODEG.	ALEXANDER, M & LUSTIGMAN, BK (1966)
100-02-7	C6H5N03	4-NITROPHENOL	BIODEG.	BEVERIDGE, EG & TALL, D (1969)

[a] The 14 data types which are used include water solubility, octanol/water partition coefficient, vapor pressure, UV spectra, dissociation constant, adsorption, bioconcentration, evaporation, Henry's law constant, biodegradation, hydrolysis, photooxidation, monitoring, and ecosystems.

16, 19, 59, 60, 61, 67, 74, 80). Table II lists the bibliographic data bases that have been examined and the ones that are most useful for obtaining pertinent data. The data bases other than Chemical Abstracts are searched by chemical name, with application of a textual search strategy when needed. NTIS and Chemical Abstracts have been found to be the most useful data bases and are always searched. Other data bases may be searched, depending on the class of chemicals and the success of the NTIS and Chemical Abstracts search. The effectiveness of the search strategy is checked by examining the references cited in relevant articles that have been retrieved. Older references are identified in this way in addition to manual searches of Chemical Abstracts.

The textual search strategy does not currently contain any terms for physical or chemical properties such as water solubility, vapor pressure, etc. Rarely did those terms provide any relevant citations, and the cost of searching was prohibitive. In addition, an attempt is made to avoid duplicating the effort necessary to retrieve data that are already available in other data bases or compilations and where the quality of the data often has been evaluated. When data are available for the subject chemicals in these other data bases, the information is directly entered into DATALOG and CHEMFATE. These data bases cover octanol–water partition coefficient,[13] water solubility,[14] vapor pressure,[15] and dissociation constants,[16–18] as well as other tabulated data.[10,19] In general, the primary search strategies are focused at obtaining environmental degradation and transport information, and physical and chemical properties are entered from secondary sources such as the data bases referenced above.

Under the present system, searching literature and adding information to the Environmental Fate Data Base involves the following steps which are discussed in more detail in the following sections:

(1) Search the primary literature for a group of chemicals (anywhere from 10 to 30) and obtain relevant abstracts, papers, and documents.
(2) For each relevant article, enter into DATALOG the abbreviated reference, data types, and CAS registry number for all chemicals examined in the article. Enter the full citation into XREF.
(3) When ready to enter data on a particular chemical into CHEMFATE, obtain a listing for the chemical from DATALOG. This listing will contain all the references obtained from the literature search as well as the references uncovered during other searches.
(4) Compare the DATALOG listing with the CHEMFATE log to see whether data from any of the references have already been entered into CHEMFATE. Excluding these, the references in the DATALOG listing are obtained from the files.
(5) Abstract relevant data from the remaining references and enter them into CHEMFATE.

DATALOG AND XREF FILES

Usually papers identified and obtained because of their relevance to one of the 200 specified chemicals also contain pertinent data on additional chemicals. Unless the articles are indexed in a way that captures the identity of these other chemicals, the information on these chemicals would have to be retrieved again at a later date and might not be retrieved even through a primary search on one of those chemicals. Primary searches on these additional chemicals would also result in duplicating articles and reports already in the files. The DATALOG file was thus created to maximize the retention of relevant data and to organize the data by chemical.

DATALOG currently contain over 21 000 records of over 2600 chemicals. The records are keyed by CAS number and contain one of 14 data types (Table III), an abbreviated reference, a common name, and chemical formula in Hill notation. The last two items are to aid in identification. The data types are defined generically to facilitate the categorization and rapid processing of references. For example, no distinction is made between biodegradation by pure cultures or mixed natural cultures; they both fall under the term "BIODEG". As papers are located and retrieved, a unique abbreviated reference and the names of the chemicals and types of data contained in the reference are written down. The CAS number is looked up, and then this number, data type, and abbreviated reference are key punched on a computer card. A batch program is used to add several hundred of these cards at a time to the DATALOG file. The record is given a key on the basis of its CAS registry number and a consecutive index number. Later entries have higher numbers, and this fact is utilized in an update program which retrieves only records added after a given date. Whenever new chemicals are added to the file, cards containing the CAS registry number, name, and chemical formula must be prepared. After initial entry, this information is retrieved by the computer when new records for the CAS registry number are enetered into the file. Full reference citations are entered into a bibliographic file, XREF, which is linked to DATALOG and CHEMFATE through the abbreviated reference. Journal articles and other references are stored alphabetically in XREF, which currently contains 1900 references.

Information can be obtained from DATALOG/XREF interactively at a computer terminal or by batch processing. In the interactive mode, the program can print out the DATALOG entries for a given CAS number with or without the full reference. The data output can be further limited by specifying any one of the 14 data types. For example, all biodegradation data for any chemical in the system can be retrieved with a full bibliography sorted alphabetically by author. DATALOG can be used in an update mode by looking up the highest index number associated with a particular CAS registry number prior to the update date. In addition, an abbreviated reference can be entered to retrieve the full citation. To aid in using the system, the program has built in "help" features. An example of an interactive session and output appears in Figure 2.

In batch mode the program has additional capabilities. The entire file can be printed, sorted by CAS number, chemical name, or chemical formula, and the file can be searched by chemical name, formula, reference, or data type. This type of search is much slower than the interactive search because

```
1XEQ RNDATALG

 RNDATALG RETRIEVES INFORMATION FROM THE
 SYRACUSE RESEARCH CORPORATION DATALOG & XREF FILES

 TYPE HELP FOR HELP      TYPE END TO END

 ENTER SEARCH TYPE
?HELP
 ENTER "SEARCH" FOR HELP ON SEARCH TYPES
 ENTER "DATA  " FOR HELP ON DATA TYPES
 ENTER "INPUT " FOR HELP ON INPUT TYPES
?SEARCH
 THE AVAILABLE SEARCH TYPES ARE:
 "CAS"  RETRIEVES DATA FOR THE PARTICULAR CAS
 "MULT" RETRIEVES DATA FOR THE PARTICULAR CAS AND TYPE OF DATA
 "CREF" RETRIEVES DATA AS IN CAS BUT ADDS COMPLETE REFERENCE
 "MREF" RETRIEVES DATA AS IN MULT BUT ADDS COMPLETE REFERENCE
 "XREF" RETRIEVES FULL REFERENCE ONLY FOR THE CITED REFERENCE

 ENTER SEARCH TYPE
HELP
 ENTER "SEARCH" FOR HELP ON SEARCH TYPES
 ENTER "DATA  " FOR HELP ON DATA TYPES
 ENTER "INPUT " FOR HELP ON INPUT TYPES
?DATA
 DATA TYPES ARE:  VP  BIODEG  WATER SOL  HENRY CON  EVAP  ECOS
   MONIT  ADSORP  UV  BIOCON  O/W PART  PHOTOOXID  DISS CONS HYDROL

 ENTER SEARCH TYPE
HELP
 ENTER "SEARCH" FOR HELP ON SEARCH TYPES
 ENTER "DATA  " FOR HELP ON DATA TYPES
 ENTER "INPUT " FOR HELP ON INPUT TYPES
INPUT
 EXAMPLES OF INPUT ARE:
 : CAS          : MULT          : XREF
 :  108-93-0    :  90-15-3      : BERTSCH,W ET AL. (1974)
 :              : BIODEG        : XREF
 :              :               : ALEXANDER,M & ALEEM,MIH (1961)
    CREF AND MREF SAME FORMAT AS CAS AND MULT RESPECTIVELY

 ENTER SEARCH TYPE
XREF
 ENTER REFERENCE
BAKER,RD & APPLEGATE,HG (1974)
 XREF           BAKER,RD & APPLEGATE,HG (1974)

      REFERENCES
 BAKER,R.D. & APPLEGATE,H.G. (1974). EFFECT OF ULTRAVIOLET RADIATION ON THE PERSISTENCE OF PESTICIDES.$ TEX. J. SCI. 25:53-9.*

 ENTER SEARCH TYPE
?MREF
 ENTER CAS NUMBER
 XXXXX-XX-X
?75-56-9
 ENTER DATA TYPE
?PHOTOOXID
 MREF      75-56-9  PHOTOOXID

 MREF      75-56-9  PHOTOOXID

        CAS        CHEMNAME       REFERENCE

      75-56-9   PROPYLENE OXIDE     BOGYO,DA ETAL (1980)
      75-56-9   PROPYLENE OXIDE     KULEVSKY,N ETAL (1969)
      75-56-9   PROPYLENE OXIDE     PITTS,JN JR. (1979)

          REFERENCES
 BOGYO,D.A.; LANDE,S.S.; MEYLEN,W.M.; HOWARD,P.H.; AND SANTODONATO,J. (1980). INVESTIGATION OF SELECTED ENVIRONMENTAL CONTAMINANTS:  EPOXIDES.$
      EPA-560/11-80-005.*
 KULEVSKY,N.; WANG,C.T.; AND STEMBERG,V.I. (1969). PHOTOCHEMICAL OXIDATION. II. RATE AND PRODUCT FORMATION STUDIES ON THE PHOTOCHEMICAL OXIDATION
      OF ETHERS.$ J. ORG. CHEM. 34:1345-8.*
 PITTS,J.N. JR. (1979). CHEMICAL CONSEQUENCES OF AIR QUALITY STANDARDS AND OF CONTROL IMPLEMENTATION PROGRAMS: ROLES OF HYDROCARBONS, OXIDES OF
      SULFUR AND AGED SMOG IN THE PRODUCTION OF PHOTOCHEMICAL OXIDANT AND AEROSOL.$ RIVERSIDE, CA: STATEWIDE AIR POLLUTION RESEARCH CENTER.*

 ENTER SEARCH TYPE
?CAS
 ENTER CAS NUMBER
 XXXXX-XX-X
?75-52-5
 CAS        75-52-5

 CAS        75-52-5

        CAS       FORM         CHEMNAME

      75-52-5   CH3NO2          NITROMETHANE

 DATA            REFERENCE

 BIODEG          MARION,CV & MALANEY,GW (1963)
 HENRY CON       BOCEK,K (1976)
 PHOTOOXID       DILLING,WL ET.AL. (1976)
 PHOTOOXID       GRAEDEL,TE (1978)
 PHOTOOXID       HAMPSON,RF JR. & GARVIN,D (1977)
 PHOTOOXID       PITTS,JN JR. (1979)
```

Figure 2. Interactive Session with DATALOG and XREF.

the file is not keyed by these items.

CHEMFATE

The CHEMFATE file contains experimental values on pertinent fate phenomena. It is divided into Chemical Identification Data, Chemodynamic Properties, Transport Properties, Degradation Studies, Field Studies, and Ambient Monitoring. Table IV lists the types and organization of data contained in CHEMFATE. The data elements were made to be compatible with the SPHERE (Scientific Parameters for Health and the Environment, Retrieval, and Estimation)

```
 !XEQ RNCHMFTE
 09:53 MAY 12, 81  PROGRAM FOR CHEMFATE
 LAST ACCOUNTING GROUP IS AT:  837 TO CHANGE ENTER NO.
?
 ENTER CAS NUMBER
?95476
 CAS NO. = 95476
   ENTER DATA TYPE
?OXID
 TYPE = OXID
 ENTER SUMMARY VALUE FOR OXID (A)
?
 ENTER OXIDANT NUMBER 1 (A;DEF=NONE)
?OH
 HOW MANY RATES AS A F(T) FOR OXIDANT  NO. 1 ARE YOU ENTERING
?1
 ENTER 1 RATES AS F(T) (ST;DEF=CC/MCULE-SEC) FOR OXIDANT NO. 1
?15.3E-12
 VALUE = .1530000E-10
 ENTER 1 HALF LIVES(NT;DEF=DAY )FOR OXIDANT NO. 1
?1.1 & $CALC, OH CONC 4.8E5 MCULE/CC
 VALUE = 1.10000
 ENTER 1 TEMPS (NT;DEF=DEGC) FOR OXIDANT NO. 1
?25
 VALUE = 25.0000
 ENTER PRODUCT NUMBER 1 FOR OXIDANT NUMBER 1
?
 ENTER OXIDANT NUMBER 2 (A;DEF=NONE)
?
 ENTER EXPERIMENTAL CONDITIONS
?FLASH PHOTOLYSIS TO FORM OH, FLOW SYSTEM, TOTAL PRESSURE 50-600 TORR (AR), SATURATED WITH COMPOUND, CONC H2O .01/.03 TORR
 ENTER ANALYTICAL METHODS USED
?RESONANCE FLUORESCENCE
 ENTER CITATION
?HANSEN,D.A. ET AL. (1976)
 *******
 1 ADD 95476 CP LDEG 838
 2 OXID A SRC
 3 SIDO 1;OH
 3 QRTO 1; .1530E-10
 3 QHL 1;1.0 & $CALC, OH CONC 4.8E5 MCULE/CC
 3 ET1 1;25.00
 3 ECON FLASH PHOTOLYSIS TO FORM OH, FLOW SYSTEM, TOTAL PRESSURE 540-600 TORR (AR), SATURATED WITH COMPOUND, CONC H2O .01/.03 TORR
 3 EMTH RESONANCE FLUORESCENCE
 3 RLIT HANSEN,D.A. ET AL. (1976)
 3 SAVE 838
 *******
  ENTER CAS NUMBER
 ?106423
  CAS NO. = 106423
    ENTER DATA TYPE
 ?/
  TYPE = OXID
  ENTER SUMMARY VALUE FOR OXID (A)
 ?
  ENTER OXIDANT NUMBER 1 (A;DEF=NONE)
 ?/
  HOW MANY RATES AS A F(T) FOR OXIDANT NO. 1 ARE YOU ENTERING
 ?1
  ENTER 1 RATES AS F(T) (ST;DEF=CC/MCULE-SEC) FOR OXIDANT NO. 1
 ?12.2E-12
  VALUE = .1220000E-10
  ENTER 1 HALF LIVES(NT;DEF=DAY ) FOR OXIDANT NO. 1
 ?1.4 & $CALC, OH CONC 4.8E5 MCULE/CC
  VALUE = 1.40000
  ENTER 1 TEMPS (NT;DEF=DEGC) FOR OXIDANT NO. 1
 ?25
  VALUE = 25.0000
  ENTER PRODUCT NUMBER 1 FOR OXIDANT NUMBER 1
 ?
  ENTER OXIDANT NUMBER 2 (A;DEF=NONE)
 ?
  ENTER EXPERIMENTAL CONDITIONS
 ?/
  ENTER ANALYTICAL METHODS USED
 ?/
  ENTER CITATION
 ?/
  *******
  1 ADD 106423 CP LDEG 839
  2 OXID A SRC
  3 SIDO 1;OH
  3 QRTO 1; .1220E-10
  3 QHL 1;1.40 & $CALC, OH CONC 4.8E5 MCULE/CC
  3 ET1 1;25.00
  3 ECON FLASH PHOTOLYSIS TO FORM OH, FLOW SYSTEM, TOTAL PRESSURE 50-600 TORR (AR), SATURATED WITH COMPOUND, CONCC H2O .01/.03 TORR
  3 EMTH RESONANCE FLORESCENCE
  3 RLIT HANSEN,D.A. ET AL. (1976)
  3 SAVE 839
  *******
  ENTER CAS NUMBER
 ?F/
  CAS NO. = 1064523
    ENTER DATA TYPE
 ?F/
  LAST ACCOUNTING GROUP IS AT: 839
  *STOP* 0
  ***XEQ TERMINATED***
```

Figure 3. Example of data input into CHEMFATE.

formats that are planned for development by the U.S. EPA Office of Pesticides and Toxic Substances.[20] The SPHERE system concept evolved from an approach originally developed for the HEEDA (Health and Environmental Effects Data Analysis) system.[21]

Each record in CHEMFATE contains at least a CAS number, data type (e.g., vapor pressure, soil thin-layer chromatography), and a reference as the key elements for identifying the chemical, the type of data contained in the record, and the citation. The record may also contain a summary value and an assortment of associated data qualifiers. Data lines are of variable length and may contain comments when

Table IV. Categories and Data Type Codes for CHEMFATE

description	code format
identification	ID
molecular formula	MF
molecular weight	MW
proper name	PNAME
synonym	SYN
chemical property	CP
chemical dynamic property	CDYN
log octanol/water partition function	LOGP
log acid dissociation constant	PKA
soil adsorption constant	SOIA
ultraviolet absorption	UV
vapor pressure	VP
water solubility	WSOL
transport properties	TRAN
log bioconcentration factor	BIOC
evaporation from water-T1/2	EVAW
Henry's law constant	HENL
soil column transport	SCOL
soil thin-layer chromatography	SRF
laboratory degradation	LDEG
ecosystem	ECOS
hydrolysis	HYDR
microbial degradation	MICD
degradation in natural system	NSYD
oxidation and other reactions	OXID
photolysis	PHOT
environmental measurement	EM
field studies and monitoring	FSMO
air monitoring	AIRM
biota monitoring	BIOM
field studies	FIEL
soil monitoring	SOIM
water monitoring	WATM

clarifying information is added. A unique set of attributes defines each data element by designating (1) the structure of the data contained therein (e.g., simple decimal, scientific, tabular, numeric range, alphanumeric, etc.), (2) the form of the default units (units that will be used unless otherwise stated), and (3) the presence of multiple data entries (repeated data for a series of species, locations, dates, etc.). The attributes are stored in a dictionary. This approach to file organization makes efficient use of computer space, facilitates data entry, and helps to standardize data entry and display.

The data elements for the 22 data subtypes were developed after careful examination of the types of fate data available in the literature and consideration of the relevant subsidiary information typically found in the articles. Subsidiary information is included either as a unique data element having its own qualifier codes or as a comment. The advantages of having the subsidiary information with its own code as a data element is that this information takes on a more dominant position in a report, and the specific data element can be searched for and extracted from the file for the purpose of editing or reporting. This is not the case with comments. The number of qualifiers for the data types ranges from one (reference only), for log of the octanol/water partition coefficient (experimental temperature is not usually reported), to nine for photolysis. For photolysis, the qualifiers relate to rate, half-life, wavelength quantum yield, product name, product CAS number, experimental conditions, and analytical method, in addition to the reference.

The data element for the reference contains the same abbreviated reference that is used in DATALOG. Just as with DATALOG, this abbreviated reference can be used to extract the full citation from the XREF file.

Data are entered into the CHEMFATE file interactively from a computer terminal. In response to the computer queries, the CAS registry number and the data type are entered. From then on, computer input prompts depend on the data type; the computer provides prompts for the relevant data,

```
**********
                    VINYL CHLORIDE              CAS#  75014
PREFERRED NAME (9CI):       CHLOROETHENE
SYNONYMS:                   VINYL CHLORIDE; CHLOROETHYLENE
MF:                         C2H3CL
MW:                         62.50
**********
                    TETRAHYDROFURAN             CAS#  109999
LOG O/W PARTITION FUNCTION:     .46
    REF:                        HANSCH,C. & LEO,A.J. (1979).
**********
                    METHYL PARATHION            CAS#  298000
SOIL ADSORPTION CONST.:     .1843E+02
    SOIL ID 1:              4 SOILS
    FREUNDLICH'S K:         .1843E+02
    1/M:                    1.04
    KOC:                    .9799E+04
    REF:                    REINBOLD,K.A. ET AL. (1979).
**********
                    N-HEXANE                    CAS#  110543
VAPOR PRESSURE (25 DEGC):   .1520E+03 TORR
    VP:                     .5900E+02           .9700E+02       .2310E+03       TORR
    TEMP:                   5.0                 15.0            35.0            DEGC
    REF:                    IWOLINSKI,B.J. & WILHOIT, R.C. (1971).
**********
                    PHENOL                      CAS#  108952
LOG BIOCON. FACTOR:         4.20
    SPECIES 1:              FATHEAD MINNOW
    LOG BCF 1:              4.20            #1
    COMMENTS:               #1 WHOLE BODY
    REF:                    CALL,D.J. ET AL. (1980).
**********
                    2-CHLOROPHENOL              CAS#  95578
MICROBIAL DEGRADATION:          ---
    SPECIES 1:              NOCARDIA BR1
    RATE 1:                 *25 CO2 IN 1 WK
    SPECIES 2:              PSEUDOMONAS B6/8, B49, B4/3
    RATE 2:                 35-85% CO2 IN 1 WK
    SPECIES 3:              PSEUDOMONAS PUTIDA
    RATE 3:                 18% CO2 IN 1 WK
    SPECIES 4:              ARTHROBACTER B5, B7
    RATE 4:                 .45 CO2 IN 1 WK
    CONDITIONS:             MICROBES FROM 9 TYPES OF SOIL, CELL SUSPENSION, 20 UG/ML, CELLS
                            GROWN ON BENZENE OR P-DIHYDROXYBENZENE
    METHOD:                 14-CO2 EVOL
    REF:                    HAIDER,K. ET AL. (1974).
**********
                    VINYL ACETATE               CAS#  108054
DEGRADATION IN NATURAL SYSTEMS:     DEGR
    SYSTEM:                 WATER
    RATE:                   34% BODT IN 5 DAY, 32% IN 20 DAY; 51% IN 5 DAY, 58% IN 20 DAY
                            (SALT WATER)
    CONDITIONS:             FILT SEW FEED
    METHOD:                 BOD
    REF:                    PRICE,K.S. ET AL. (1974).
**********
                    BENZENE                     CAS#  71432
AIR MONITORING:             14.00 UG/CU-M, *AV, 3 U.S. CITIES
    DATE 1:                 SUMMER-FALL 1977
    SITE 1:                 CHICAGO
    SITE 2:                 DALLAS
    SITE 3:                 LOS ANGELES
    CONC (DATE I, SITE J):  DATE                1
                            SITE 1              18.00           UG/CU-M         #1
                            SITE 2              5.00            UG/CU-M         #2
                            SITE 3              19.00           UG/CU-M         #3
    SAMPLE TYPE:            SAMPLE ABSORBED ON TENEX RESIN BEADS IN SAMPLING CARTRIDGE
    METHOD:                 GC-FID
    COMMENTS:               #1 AV, INCL. 9 UG/CU-M CORRECTION DUE TO RESIDUAL BENZENE IN TENEX
                            #2 AV, INCL. 9 UG/CU-M CORRECTION DUE TO RESIDUAL BENZENE IN TENEX
                            #3 AV, INCL. 9 UG/CU-M CORRECTION DUE TO RESIDUAL BENZENE IN TENEX
    REF:                    MARTIN,B.E. ET AL. (1980).
**********
                    DI(2-ETHYLHEXYL) PHTHALATE  CAS#  117817
ECOSYSTEM:                  ---
    EXP 1:                  TERRESTRIAL - AQUATIC SYSTEM
    COMPONENT 1:            WATER
    COMPONENT 2:            ALGAE (OEDOGONIUM)
    COMPONENT 3:            SNAIL (PHYSA)
    COMPONENT 4:            MOSQUITO (CULEX)
    COMPONENT 5:            GUPPY (GAMBUSIA)
    CONC (EXP I, CMPT J):   EXP                 1
                            CMPT. 1             .0078 PPM (TOTAL C-14); .00034 PPM (DEHP)
                            CMPT. 2             19.105 PPM (TOTAL C-14); 18.322 PPM (DEHP)
                            CMPT. 3             20.325 PPM (TOTAL C-14); 7.302 PPM (DEHP)
                            CMPT. 4             36.509 PPM (TOTAL C-14); 36.509 PPM (DEHP)
                            CMPT. 5             .206 PPM (TOTAL C-14); .044 PPM (DEHP)
    LOG BIOCONC.(EXP I, CMPT J): EXP            1
                            CMPT. 1             ---
                            CMPT. 2             4.73 (DEHP)
                            CMPT. 3             4.33 (DEHP)
                            CMPT. 4             5.03 (DEHP)
                            CMPT. 5             2.11 (DEHP)
    CONDITIONS:             5 MG DEHP APPLIED TO TERRESTRIAL PART OF ECOSYSTEM, 33 DAYS, STATIC TEST.
    METHOD:                 C-14, TLC
    REF:                    METCALF,R.L. ET AL. (1973).
**********
```

Figure 4. Examples of CHEMFATE records.

entry format, and default units. After each entry, the computer codes, formats, and enters the data into CHEMFATE. The interactive program has jumps and loops built to accommodate the expected variety of responses. There is a provision for repeating an entry of a particular data element. This increases the efficiency of entering data on several subject chemicals from a single article. A record is completed after the abbreviated reference is entered. At this time a consecutive record number is assigned to the entry. Hence, CHEMFATE is continuously updated as data is entered at the terminal. A ledger (file named ACCT) which includes the CAS number, data type, reference, record number, and entry date provides basic file maintainance and is automatically updated as new records are transferred to CHEMFATE. Two examples of data input to CHEMFATE are provided in Figure 3.

Quality control of the CHEMFATE data base is assured in several ways. By having an interactive program for data entry where the computer asks questions and formats the response in the file, errors in codes for data type and qualifiers are virtually eliminated. In addition, when a CAS registry number is requested, a check is automatically performed (using

the algorithm on the check digit) to see whether the number is a valid one, thereby eliminating most keying errors for registry numbers. Similarly, alphanumeric strings will not be accepted when numerical data are required. After a record is entered at a terminal, a copy of the record as it will be entered into the CHEMFATE file and the record number is displayed on the screen, giving the person entering the data an opportunity to review what has been entered and to make notes on what should be edited. After these corrections are made, a copy of the records are returned to the individual who abstracted the information to recheck the file entry.

The CHEMFATE records have a two-faceted key. The first part is the record number, and the second part is an interrecord line number. CHEMFATE records can be retrieved rapidly by searching ACCT which is keyed by CAS registry number for the desired registry number, obtaining the relevant record numbers and retrieving these records from CHEMFATE.

The report generation program for CHEMFATE has been completed. For a specified CAS registry number this program retrieves and formats all records on the chemical, any of the 22 data types or ID (identification) (see Table IV), or the four groups of data types, namely, chemical dynamic properties, transport properties, laboratory degradation, and field studies and monitoring. Examples of CHEMFATE records appear in Figure 4. It is planned to make both DATALOG and CHEMFATE data bases available to the public.

ACKNOWLEDGMENT

The support for this work from the U.S. Environmental Protection Agency, Office of Toxic Substances, under Cooperative Agreement CR 806902020 is gratefully acknowledged.

REFERENCES AND NOTES

(1) Conway, R. A. "Environmental Risk Analysis of Chemicals"; Van Nostrand Reinhold: New York, 1981.
(2) Howard, P. H.; Santodonato, J.; Durkin, P. R. "Syracuse Research Corporations Approach to Chemical Hazard Assessment". In "Environmental Risk Analysis for Chemicals"; Van Nostrand Reinhold: New York, 1981.
(3) Christensen, H. E.; Fairchild, E. J.; Carroll, B. S.; Lewis, R. L., Sr. "Registry of Toxic Effects of Chemical Substances"; NIOSH, U.S. Government Printing Office: Washington, DC, 1980.
(4) Christensen, H. E.; Fairchild, E. J.; Lewis, R. L. "Suspected Carcinogens"; 2nd ed.; NIOSH, U.S. Government Printing Office: Washington, DC, 1976.
(5) Kemp, H. T.; Little, R. L.; Holoman, V. L.; Darby, R. L. "Water Quality Criteria Data Book, Vol. 3 and 5. Effects of Chemicals on Aquatic Life"; U.S. EPA, U.S. Government Printing Office: Washington, DC, 1971, 1973.
(6) Fishbein, L. "Potential Industrial Carcinogens and Mutagens"; U.S. EPA 560/5-77-005.
(7) National Library of Medicine. "Toxicity Data Bank"; U.S. Department of Health, and Human Services: Washington, DC.
(8) Magnuson, V.; Harriss, D.; Maanun, W.; Fulton M. "ISHOW User's Manual"; University of Minnesota: Duluth, MN, 1979.
(9) OHMTADS (Oil and Hazardous Materials Technical Assistance Data System), NIH/EPA Chemical Information System.
(10) For example, WATERDROP (Distribution Register of Organic Pollutants NIH/EPA Chemical Information System) and STORET (STOrage and RETrieval of water quality data), U.S. EPA Office of Water and Hazardous Materials, Washington, DC.
(11) Howard, P. H.; Saxena, J.; Sikka, H. C. "Determining the Fate of Chemicals". *Environ. Sci. Technol.* **1978**, *12*, 398-407.
(12) SRI International. "A Study of Industrial Data on Candidate Chemicals for Testing"; EPA 560/5-77-006; U.S. Nat. Tech. Inform. Serv. PB274264.
(13) Hansch, C.; Leo, A. J. "Substituent Constants for Correlation Analysis in Chemistry and Biology"; New York, 1979.
(14) S. Yalkowsky at Upjohn in Kalamazoo, MI, has collected aqueous solubility data on 2000 nonelectrolyte solutes.
(15) Zwolinski, B. J. and Wilhoit, R. C. "Handbook of Vapor Pressure and Heats of Vaporization of Hydrocarbons and Related Compounds"; Thermodynamics Research Center: College Station, TX, 1971; API44-TRC101.
(16) Perrin, D. D. "Dissociation Constants of Organic Bases in Aqueous Solution"; Buttersworth: London, 1965; IUPAC Chemical Data Series.
(17) Perrin, D. D. "Dissociation Constants of Organic Bases in Aqueous Solution"; Buttersworth: London, 1972; IUPAC Chemical Data Bases: Supplement 1972.
(18) Serjeant, E. P.; Dempsey, B. "Ionisation Constants of Organic Acids in Aqueous Solution"; Pergamon Press: New York, 1979; IUPAC Chemical Data Series.
(19) Hampson, R. F. "Chemical Kinetics and Photochemical Data Sheets for Atmospheric Reactions"; U.S. Department of Transportation: Washington, DC, 1980; Report FAA-EE-80-17.
(20) "Scientific Parameters in Health and the Environment, Retrieval and Estimation: A Requirement Analysis and Examination of Alteratives"; CRC Systems Incorporated: Fairfax, VA, 1981; EPA Contract 68-01-4795.
(21) Lefkovitz, D.; Rispin, A.; Kulp, C.; Hill, H. "EPA Health and Environmental Effects Data Analysis System". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 18-28.

Fast, Parallel Relaxation Screening for Chemical Patent Data-Base Search

LES KITCHEN and E. V. KRISHNAMURTHY*

Computer Vision Laboratory, Computer Science Center, University of Maryland, College Park, Maryland 20742

Received September 14, 1981

Described here is an application of the discrete relaxation scheme to search for specific structures and substructures which are included within the generic chemical structure expressions in a chemical patent data base. This scheme can be made highly parallel, since only the local compatibility conditions that are independent are checked, and these checks can be performed simultaneously on a parallel multiprocessor computer, with enormous savings in computation time.

INTRODUCTION

One of the greatest problems encountered in dealing with the information in chemical patents is the widespread use of generic chemical nomenclature or Markush expressions, where classes of molecules are described which may be either finite or potentially infinite in number, depending upon the constraints placed on the possible position and variety of substituents or other variable characteristics. The economic importance of constructing an efficient computer-based information system for this purpose is now clearly understood.

Most current chemical information systems are widely used for the retrieval of specific structures and of groups of compounds related by their having substructures in common and so are as such inadequate to handle generic structural information. Also, these use essentially node-by-node sequential

* Indian Institute of Science, Bangalore, India.

0095-2338/82/1622-0044$01.25/0