TERRE-TOX: A Data Base for Effects of Anthropogenic Substances on Terrestrial Animals

S. MARK MEYERS* and SUSAN M. SCHILLER

Northrop Services, Inc., Corvallis, Oregon 97333

Received July 22, 1985

TERRE-TOX is a new data base developed for the U.S. Environmental Protection Agency to aid in evaluating premanufacturing notices and research. TERRE-TOX contains published (1970 to present) information on toxicity of anthropogenic substances to terrestrial animals. Currently, species are limited to wildlife, bees, earthworms, and laboratory rodents where the substance involved is likely to affect wildlife. When records from the Denver Wildlife Research Center are incorporated, there will be approximately 15000 studies dealing with acute toxicity, behavior, reproduction, physiological, and biochemical responses. Additional information on the chemical used for each study (CAS Registry Number, purity, synonyms), test organisms (species, common name, age), and test conditions (route of administration, dose regimen, range of doses) is also provided. TERRE-TOX is designed to become part of SPHERE (Scientific Parameters for Health and Environment, Retrieval and Estimation) within the Chemical Information System (CIS). In addition, capabilities for research into structure–activity relationships (SAR) will be possible after link-up with SANSS (Structure and Nomenclature Search System).

INTRODUCTION

Large computerized data base systems can provide researchers and regulatory agencies with rapid dissemination and analysis capabilities of data otherwise available only through extensive manual review of published literature. The Environmental Protection Agency (EPA) recognizes the need for computerized data base systems for risk assessment and has developed comprehensive systems like Health and Environmental Effects Data Analysis (HEEDA)[1] and Scientific Parameters for Health and the Environment, Retrieval and Estimation (SPHERE).[2] These systems, developed for the EPA's Office of Toxic Substances (OTS), were designed to aid staff in risk assessment relative to premanufacturing notice reviews (PMN) by providing a controlled information storage and retrieval system that would allow on-line users to analyze for structure–activity relationships (SAR) and comparative toxicity.[2]

Our goal for developing a terrestrial animal toxicity data base (TERRE-TOX) was to provide a compilation of published research data that would become a subcomponent of SPHERE. Data as well as journal citations would be retrieved through a number of control categories, including Chemical Abstracts Service Registry Numbers, chemical names, species, test characteristics, and test results.

As a component of SPHERE, TERRE-TOX would parallel existing data bases dealing with environmental toxicology that are already a part of the system or are scheduled for inclusion. Among these data bases are PHYTOTOX[3] (chemical effects on terrestrial plants), AQUIRE[4] (chemical effects on aquatic organisms), and CHEMFATE[5] (fate of chemicals in the environment).

TERRE-TOX is not the first compilation of toxicity data on terrestrial animals (wildlife). Several references are available containing tabulations of toxicity data,[6–9] some of which are available on-line.[10] There remains a large body of information dealing with a wide spectrum of data types (study purposes, effective end points) and chemicals in hundreds of internationally published journals that until this time have not been readily available. We have attempted to collect and process all pertinent data published since 1970.

DATA BASE DESCRIPTION

The TERRE-TOX data base programs were written with the INFORM-11[11] data base management system of the U.S. Environmental Protection Agency's research laboratory, in Corvallis, OR. INFORM is a large program used primarily for data storage, retrieval, and manipulation. It has some analytical capabilities (descriptive statistics, regression, polynomial curve generation). The screens for the TERRE-TOX data base were designed in INFORM to be accessed on a Digital VT50 or VT100 terminal or on a terminal that has the capacity to emulate the VT100.

TERRE-TOX is comprised of a bibliographic file and a date file. The bibliographic file contains 4600 citations of published articles from approximately 480 journals. The data file is made up of information that has been extracted from the articles in the bibliographic file. It consists of test data of more than 150 species and 800 chemicals. At present, there is a total of 3200 studies entered into the data file.

Bibliographic File. The Bibliographic File was created initially as a listing of all articles collected, reviewed, and accepted for TERRE-TOX, and which would eventually be coded for the data file. Articles on the effects of anthropogenic substances on birds, mammals, reptiles, terrestrial stages of amphibians, and selected invertebrates were collected. This includes laboratory, pen, and field studies. The bibliographic file can be accessed on-line and can be searched by reference identification number, author, title (word search), journal, citations, year, or category, with hard-copy capabilities. In addition, the entire bibliographic file can be printed alphabetically, by author. It is useful as a source for quick title searches on recent animal toxicology literature. It also connects the information on the data file with its appropriate citation for easy identification of the published articles.

To locate the greatest percentage of available literature appropriate to TERRE-TOX, three search techniques were utilized. Searches were conducted manually (hand searching), via computerized literature data bases, or by searching the bibliographies of articles already in TERRE-TOX (tree searching).

Hand searching of up to 14 specific journal titles for any given year provided 100% of the appropriate literature in those journals for each year (Table I). These journals were chosen with the understanding that they contained much of the literature being currently published on animal toxicology. The purpose of these hand searches was to provide a check on the accuracy of the computerized literature searches. By matching articles on the computerized searches with those located by means of hand searching, the percentage of target literature located by the computerized searches could be calculated.

At present, *Current Contents*, *Ecology Abstracts*, *Envi-*

0095-2338/86/1626-0033$01.50/0 © 1986 American Chemical Society

Table I. List of Journal Titles Used in Computerized Literature Search Verification for TERRE-TOX

Archives of Environmental Contamination and Toxicology
Bulletin of Entomological Research
Bulletin of Environmental Contamination and Toxicology
Environmental Research
Journal of Animal Science
Journal of Toxicology and Environmental Health
Journal of Wildlife Management
Marine Pollution Bulletin
New Zealand Journal of Experimental Agriculture
Pesticide Monitoring Journal
Poultry Science
Toxicology and Applied Pharmacology
Veterinary Record

ronment Abstracts, *Pollution Abstracts*, and *Wildlife Reviews* are screened regularly, and pertinent literature is collected. After commercial vendors update their files, a new computerized search is conducted to locate additional articles.

The BIOSIS Previews data base was accessed for all computerized literature searches. BIOSIS was selected because it could provide the greatest amount of appropriate literature. The use of the biosystematic and concept codes in the BIOSIS data base provided broad search capabilities. Concepts such as toxicology, wildlife management, pollution, and pesticides were searched. The use of other data bases was explored but did not prove as efficient for our purposes, as they required a more specific search strategy using, for example, chemical names and genus/species as keywords to locate the literature. It was not practical to search all conceived chemicals and species involved. It is possible that in the future, when gaps in data on specific chemical or species are recognized, specialized searches may be devised.

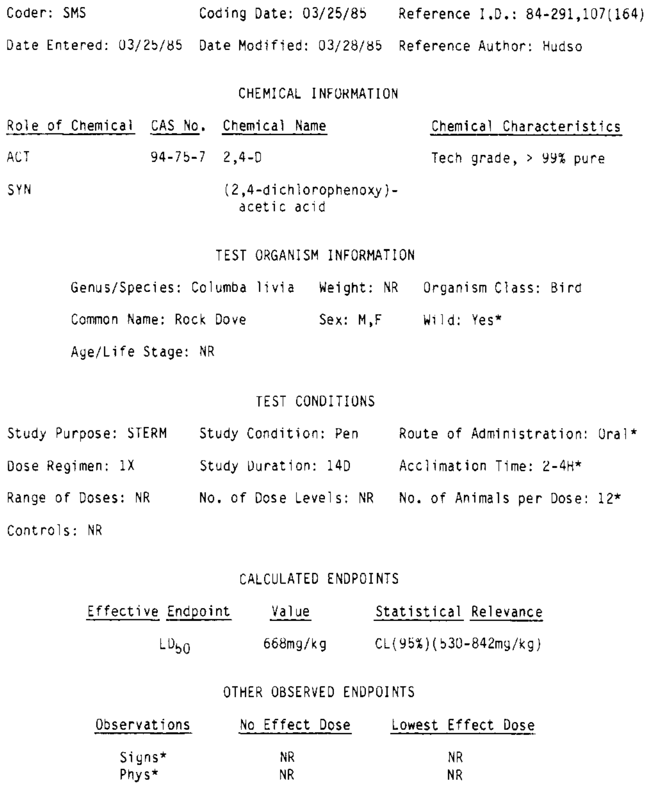
Tree searching techniques were implemented to obtain literature not located by hand or computerized searching. Tree searching involved reviewing the bibliographies of articles already acquired for TERRE-TOX. Relevant citations were reviewed, collected, and added to TERRE-TOX.

Through hand, computerized, and tree searching, we estimate that more than 90% of the literature appropriate to TERRE-TOX has been collected.

Data File. After articles had been obtained and entered into the bibliographic file, acceptable articles were coded and entered into the data file. The material in the data file includes chemical information, test organism information, test conditions, calculated end points (i.e., LD_{50}, LC_{50}), other results (signs of toxicity, behavior, avoidance, physiological, etc.), and comments. The data file can be accessed on-line and can be searched on any of the available fields. Hard-copy formats that are available include not only the data in its entirety but also various shortened versions with the specific data only.

Priorities were established for selecting articles for coding and entry. Studies on wild species and laboratory species represented in the wild had first priority for coding and have been entered into the data file. Articles on domestic animals, methods/methods development on wild species, nonefficacy insect research, heavy metals, oil, and egg injection experiments had lower priority and have not yet been coded. These articles have been entered on the bibliographic file and archived for future reference. In addition, articles involving the use of laboratory animals (rat, mouse, rabbit, dog, etc.) were collected and entered into the data file when the chemical involved had relevance to wild species.

Coding involved extracting data from each article. A data collection form for each study in the article was completed (Figure 1). When possible, data were directly entered on-line. The data file is comprised of individual studies from the articles. A new study was entered whenever there was a change in the test chemical, solvent, or vehicle or change in concentration or form of the test chemical. A new study was also entered for each new species, different age, or life stage at dosing, whenever study type or purpose changed, and for a change in the route of administration or a change in study duration.



```
Coder: SMS               Coding Date: 03/25/85   Reference I.D.: 84-291,107(164)

Date Entered: 03/25/85  Date Modified: 03/28/85  Reference Author: Hudso

                           CHEMICAL INFORMATION

Role of Chemical  CAS No.  Chemical Name               Chemical Characteristics

ACT               94-75-7  2,4-D                       Tech grade, > 99% pure

SYN                        (2,4-dichlorophenoxy)-
                             acetic acid

                         TEST ORGANISM INFORMATION

        Genus/Species: Columba livia   Weight: NR   Organism Class: Bird

        Common Name: Rock Dove         Sex: M,F     Wild: Yes*

        Age/Life Stage: NR

                              TEST CONDITIONS

Study Purpose: STERM    Study Condition: Pen      Route of Administration: Oral*

Dose Regimen: 1X        Study Duration: 14D       Acclimation Time: 2-4H*

Range of Doses: NR      No. of Dose Levels: NR    No. of Animals per Dose: 12*

Controls: NR

                            CALCULATED ENDPOINTS

            Effective Endpoint     Value       Statistical Relevance

                      LD50         668mg/kg    CL(95%)(530-842mg/kg)

                          OTHER OBSERVED ENDPOINTS

             Observations     No Effect Dose     Lowest Effect Dose

                Signs*              NR                  NR
                Phys*               NR                  NR
```

Figure 1. Terrestrial toxicity data collection form, option 1.

The information in a record of the data file is divided into six sections. The first section provides general information about the articles, such as reference I.D. number, reference author, number of tables, and number of graphs in the article.

The chemical information in a study is listed in the second section. The role of the chemical is identified by the following codes: ACT (active ingredient), TEST (test substance, can be a mixture or formulated product), SYN (synonym of the tested chemical), MF (molecular formula), INGR (ingredient of a test mixture or product, including the use of marker dyes), IMP (impurity), SOLV (solvent or vehicle), MOD (modifier, chemical that exerts a synergistic or activating effect on the tested chemical), and MET (metabolite).

In addition, the chemical information section includes the Chemical Abstracts Service (CAS) Registry Numbers, chemical names, and chemical characteristics such as purity, grade, dosage form, concentration of test solution, ratio of mixture components, and other information that characterizes the chemical.

Test organism information listed in the next section includes the genus and species, common name, and organism class. Also included are the age or life stage, weight, and sex of the organism. The organism is identified as to whether it is a wild or domestic species. The following section profiles the test conditions. The study purpose code is selected from controlled vocabulary and is used to record the author's major purpose in conducting the experiments (Table II).

Under test conditions, the study purpose, study condition, route and method of administration, dose regimen, study duration, acclimation time, range of doses, number of dose levels, and number of animals per dose are profiled. In addition, controls are designated as being concurrent (CON), base line (BASE), historic (HIST), none, or not reported (NR). Study conditions are either FIELD, PEN (outdoor enclosure or ex-

Table II. Study Purpose Codes Used in the TERRE-TOX Data Base

code	definition
ADME	absorption, distribution, metabolism, or excretion
AVOID	avoidance or repellency
BCM	biochemical
BHV/NS	behavioral and nervous system study
ECOS	ecosystem study
EFF	efficacy
FK	field kills or effects (used for anecdotal reports)
IND	indirect or secondary toxicity study
ORG	organ-specific study
OTHER	other (use comment to describe)
PHYS	physiological effects not under another category
REP	both sexes dosed in reproductive effects study
REP-F	female reproductive effects
REP-M	male reproductive effects
RES	residue
STERM[a]	short term (exposure < 8 days)
ITERM[a]	intermediate term (exposure 8–35 days)
LTERM[a]	long term (exposure > 35 days)

[a] Reserved for lethality and nonspecific toxicity studies.

Table III. Route and Method of Administration Codes Used in the TERRE-TOX Data Base

code	description
DRM	dermal
EYE	eye
INH	inhalation
INJ	injection (parenteral)
IM	intramuscular
IP	intraperitoneal
IV	intravenous
SC	subcutaneous
ORAL	includes gavage, capsule, or unspecified oral administration
DIET	diet
DW	drinking water
OTHER	specify in comments
POI	poisoning uncontrolled exposure
VITRO	in vitro

Table IV. Observation Codes Used in the TERRE-TOX Data Base

code	definition
AVOID	avoidance or repellency
BCM	biochemical
BHV	behavior
DEATH	death
GRO	growth/development
PATH	pathological
PHYS	physiological
POP	population
REP	reproduction
SIGNS	signs of toxicity (i.e., symptoms)
ABS	absorption
DIST	distribution
EXC	excretion
MET	metabolism

closure), or LAB (indoor). The route of administration of the test chemical is recorded with controlled terms (Table III).

Results of the study are divided into two groups. The first lists the calculated end points and can take a variety of forms. For example, LD50 and LC50 (median lethal dose or concentration), R50 (median repellency or avoidance dose or concentration), and ED50 and EC50 (median effective dose or concentration). The calculated value and statistical reliability are recorded for each effective end point.

Other observed end points are reported in the second group. For effects other than LD50, R50, and ED50, observations of effects and no- effects are described with more controlled vocabulary (Table IV). Corresponding to each observation, the appropriate no effect dose (N.E.D.) level and the lowest effect dose (L.E.D.) level are recorded.

A comments section concludes each data file record. Comments are entered when necessary or helpful and are used to clarify or expand data entered in a specific field. Symbols (*, #, ?) are placed at the end of fields to direct the user to the comments, related studies, or the original article.

Future for TERRE-TOX. When TERRE-TOX was initiated, it was intended for inclusion into the Chemical Information System (CIS),[12] of which SPHERE is a subcomponent. At that time, CIS was jointly maintained by the National Institutes of Health (NIH) and EPA. Now CIS is under private ownership.[13] It is not certain at this time if TERRE-TOX will be integrated into existing systems, but interest has been expressed by the private sector.

To realize the full potential of a terrestrial toxicity data base such as TERRE-TOX, it is essential that it be incorporated into SPHERE or a similar system that could provide rapid data retrieval and have analytical capabilities. In addition, link-up with the Structure and Nomenclature Search System (SANSS),[14] also part of CIS, will add TERRE-TOX data to the current SAR capabilities.

We are currently attempting to add to TERRE-TOX more than 10 000 data records from experimentation conducted at the Denver Wildlife Research Center.[15] The computerization of these data will be of enormous benefit to research and environmental risk assessment.

In addition, the remaining set-aside articles on domestic animals, laboratory animals, invertebrates, methods, etc. are not available in the data file. Future priorities for TERRE-TOX will also include processing these articles, thereby broadening the contents and utility of the data base.

ACKNOWLEDGMENT

The development of this data base has been funded wholly by an Environmental Protection Agency contract (68-03-3246) to Northrop Services, Inc. The material described in this article has been subject to the Agency's review, and it has been approved for publication as an EPA document. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

REFERENCES AND NOTES

(1) 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.

(2) "Scientific Parameters in Health and the Environment, Retrieval and Estimation: A Requirement Analysis and Examination of Alternatives"; CRC Systems Incorporated: Fairfax, VA, 1981; EPA Contract 68-01-4795.

(3) Royce, L. C.; Fletcher, J. S.; Risser, P. G.; McFarlane, J. C.; Benenati, F. E. "PHYTOTOX: A Data Base Dealing with the Effect of Organic Chemicals on Terrestrial Vascular Plants". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 7–10.

(4) "Aquatic Information Retrieval Data Base (AQUIRE): Project Description, Guidelines, and Procedures"; Environmental Research Laboratory, U.S. Environmental Protection Agency: Duluth, MN, 1981.

(5) Howard, P. H.; Sage, G. W.; Lamacchia, A. "The Development of an Environmental Fate Data Base". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 38–44.

(6) Heath, R. G.; Spann, J. W.; Hill, E. F.; Kreitzer, J. F. "Comparative Dietary Toxicities of Pesticides to Birds". *Spec. Sci. Rep.: Wildl.-U.S., Fish Wildl. Serv.* **1972**, *No. 152.*

(7) Hill, E. F.; Heath, R. G.; Spann, J. W.; Williams, J. D. "Lethal Dietary Toxicities of Environmental Pollutants to Birds". *Spec. Sci. Rep.: Wildl.-U.S., Fish Wild. Ser.* **1982**, *No. 191.*

(8) Schafer, E. W. "The Acute Oral Toxicity of 369 Pesticidal, Pharmaceutical, and Other Chemicals to Wild Birds". *Toxicol. Appl. Pharmacol.* **1972**, *21*, 315–330.

(9) Schafer, E. W., Jr.; Bowles, W. A., Jr.; Hurlbut, J. "The Acute Oral Toxicity, Repellency, and Hazard Potential of 998 Chemicals to One or More Species of Wild and Domestic Birds". *Arch. Environ. Contamin. Toxicol.* **1983**, *12*, 355–382.

(10) Walker, C. R.; Menzie, C. M.; Bowles, W. A., Jr. "Evaluation of an Information Retrieval System for Assessment of Toxicological Effects of Chemicals on Fish, Wildlife, and Ecosystem Components". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 29–35.

(11) "INFORM-11 Reference Manual"; United Computing Systems, Inc.: Kansas City, MO, 1977.

(12) Milne, G. W. A.; Heller, S. R. "NIH/EPA Chemical Information

System". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 204–211.

(13) We are aware of two private CIS vendors: Fein-Marquart Associates, Baltimore, MD, and ICI, Inc., Washington, DC. Files for CIS are also available from the National Technical Information Service (NTIS).

(14) Milne, G. W. A.; Heller, S. R.; Fein, A. E.; Frees, E. F.; Margaret, R. E.; McGill, J. A.; Miller, J. A.; Spiers, D. S. "The NIH-EPA Structure and Nomenclature Search System". *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 181–186.

(15) Schafer, E. W., Jr., Wildlife Research Center, U.S. and Wildlife Service, Denver, CO, personal communication.

Implementation of Nearest-Neighbor Searching in an Online Chemical Structure Search System

PETER WILLETT* and VIVIENNE WINTERMAN

Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, U.K.

DAVID BAWDEN

Research Information Services, Pfizer Central Research, Sandwich, Kent CT13 9NJ, U.K.

Received August 6, 1985

This paper discusses the provision of nearest-neighbor searching facilities as an adjunct to the retrieval mechanisms of conventional chemical search systems. The facilities are based upon the calculation of a measure of intermolecular structural similarity between a query compound and the molecules in a machine-readable structure file. Examples are presented of the use of nearest-neighbor searching to rank output from substructure searches and to provide a means for carrying out browsing searches in structure files.

DETERMINATION OF INTERMOLECULAR SIMILARITY

Current systems for the retrieval of chemical structure information offer several types of search facility.[1] Registration or structure search involves the comparison of a specified query compound with each of the molecules in a file to identify an exact match, while substructural retrieval involves a partial match search for those molecules that contain the query as a substructure. A further type of search, superstructure search, may be of use in synthesis design programs.[2] A useful adjunct to such search facilities would be the provision of best match, or nearest-neighbor, routines that permitted the ranking of the compounds in a file in order of decreasing similarity to a query structure or substructure, the ranking being based upon some quantitative measure of intermolecular similarity or distance.

The concept of molecular similarity is not a new one. Thus, Adamson and Bush[3] evaluated a range of similarity measures on the basis of the fragment substructures common to a pair of compounds; Wilkins and Randic[4] and Gabanyi et al.[5] have discussed the use of simple topological relationships between compounds while Willett[6] has reported a comparison of hierarchic clustering procedures that are based on different similarity criteria. However, these studies have all been in the structure–property context, involving the use of only small sets of compounds, and until recently, there do not seem to have been any reports in the literature of the use of similarity-based ranking methods as a general search mechanism in computerized structure retrieval systems. Very recently, Carhart et al.[7] have described similarity matching procedures that are closely related to the work reported here, which discusses the implementation of nearest-neighbor searching in SOCRATES, the interactive chemical and biological data search system that has been developed at Pfizer Central Research (U.K.). The main areas of difference between our work and that of Carhart et al. are the types of structural feature and similarity measure that are used and in our adoption of an efficient best match search algorithm that is based upon the inverted file organization.

To explain the approach that we have developed for interactive best match structure searching, a brief description is required of the chemical searching component of the SOCRATES system. SOCRATES includes a chemical graphics module based upon connection table representations for each of the molecules in the file, and these tables are used for the generation of a set of fragment bit strings, one for each molecule. The strings are stored for search as a bit map, *B*, in which bit $B[I,J]$ is set to one if the *J*th fragment screen has been assigned to the *I*th molecule: currently, 1315 screens are used for the characterization of a file of over 200 000 compounds. The experiments reported here involved the use of a small subset of this file, containing the 8000 compounds in the Pfizer Stores File.

The bit map can be regarded either as a serial file, in which the bit strings are inspected in sequence one after the other, or as an inverted file, in which access is available to all of the compounds possessing a specific fragment screen. The use of the bit map in this latter form provides sufficiently fast screening for interactive substructure search by the intersection of the inverted file lists corresponding to the fragments in a query substructure; registration and structure search is carried out with the topological search codes that have been described in an earlier paper.[8]

To enable a set of compounds to be ranked in response to a query, some quantitative definition of intermolecular similarity is required,[9] and we have used the work of Adamson and Bush[3] as a basis for the systems developed here. Specifically, it is assumed that each of the structures under consideration is characterized by a set of substructural fragment descriptors and that the degree of similarity between some pair of structures or substructures can be evaluated in terms of the fragments that are, or are not, common to both of them. Given such fragment occurrence data for a pair of molecules, a very wide range of types of measure may be used to determine the degree of similarity between the two compounds. On the basis of simulated property prediction experiments using a small set of compounds with local anaesthetic activity, Adamson and

0095-2338/86/1626-0036$01.50/0 © 1986 American Chemical Society