and containing data/information on specific, definable chemical substances is required to contain the CAS Registry Number for each chemical substance.

(3) CAS developed its Chemical Registry System during the last decade through a program jointly funded by the National Science Foundation and the ACS and with early support from the Department of Health, Education and Welfare and the Department of Defense. For a general description of the CAS Chemical Registry System, see: P. G. Dittmar, R. E. Stobaugh, and C. E. Watson, "The Chemical Abstracts Service Chemical Registry System. I. General Design," in press.

The Toxicology Data Bank[†]

MICHAEL A. OXMAN*, HENRY M. KISSMAN, JOAN M. BURNSIDE, JERRY R. EDGE,
CAROL B. HABERMAN, and ARTHUR A. WYKES

Toxicology Information Program, National Library of Medicine, Bethesda, Maryland 20014

Received August 26, 1975

This paper describes the Toxicology Data Bank, a new system under development that will provide on-line access to chemical, physical, toxicologic, pharmacologic, use, and manufacturing data on 4000–5000 selected chemicals including drugs.

Over the past few years tremendous concern has been generated about the impact of numerous chemicals on the environment. As a result, various efforts are underway to compile some of the massive amounts of data that have been produced on many of these chemicals into accessible and conveniently usable forms.

Data on chemicals of interest to those involved with public health and safety are usually available from a broad spectrum of sources ranging from laboratory data sheets, technical reports, and the primary literature to monographs and textbooks. Frequently, the absence of a central repository for multidisciplinary information and data places serious constraints upon individuals seeking specific facts.

Certainly a compendium of data gathered from standard literature sources would be a significant contribution toward alleviating the current problem. However, the availability of sophisticated automated information systems and communications networks presents an entirely different realm for instantaneous access, correlation, and retrieval of selected facts from massive amounts of data. This paper describes one attempt to utilize current technology to satisfy a broad range of needs.

The National Library of Medicine (NLM), National Institutes of Health (NIH), having a mandate from the United States Congress to apply its resources broadly to the advance of the medical and health-related sciences collects, organizes, and makes available biomedical information to investigators, educators, and practitioners. As part of this mission, a project has been initiated through the Library's Toxicology Information Program (TIP) to build the Toxicology Data Bank (TDB). The purpose of the project is to meet needs in industry, government, and academia for a publicly accessible, on-line, interactive computer-based data retrieval system in toxicology. It will be the first "data" file to join the family now available through the NLM's on-line services such as MEDLINE, TOXLINE, and CHEMLINE.

It is expected that the ultimate size of the Toxicology Data Bank will be 4000–5000 chemical records. A record contains available verbal and numerical data (Figure 1) from selected sources, almost all of which are evaluated, on chemistry, physical properties, pharmacology, toxicology, manufacturing, shipping, and usage.

The basic scheme for building and maintaining the TDB is shown in Figure 2. Selected chemicals are assigned to data extraction teams. Appropriate data are encoded onto specially designed sheets and converted to machine-readable form. After final edit, the machined data are read into MARK IV, a file management system selected for building records and maintaining the data bank. From MARK IV a special report is generated in a format developed by the staff for use by a scientific review committee. This group, made up of members of the NIH's Toxicology Study Section, reviews each record for scientific content and merit. Following peer review, appropriate editing or other changes are made. Finally, the MARK IV file will be transferred to the on-line retrieval software system for public access.

The typical partial record in Figure 3 illustrates some of the file's features. Certain fields, such as Animal Toxicity, contain textual material. This material is extracted and encoded as it occurs in the literature source to eliminate the possibility of lost data or altered data appearing in the TDB. To facilitate retrieval, extracts are indexed using Medical Subject Headings (MeSH) terms, a standardized vocabulary developed by the National Library of Medicine. Retrieval, then, will be possible by searching any terms appearing in the free text or searching on selected MeSH terms. Associated with each data value is the source from which it is excerpted.

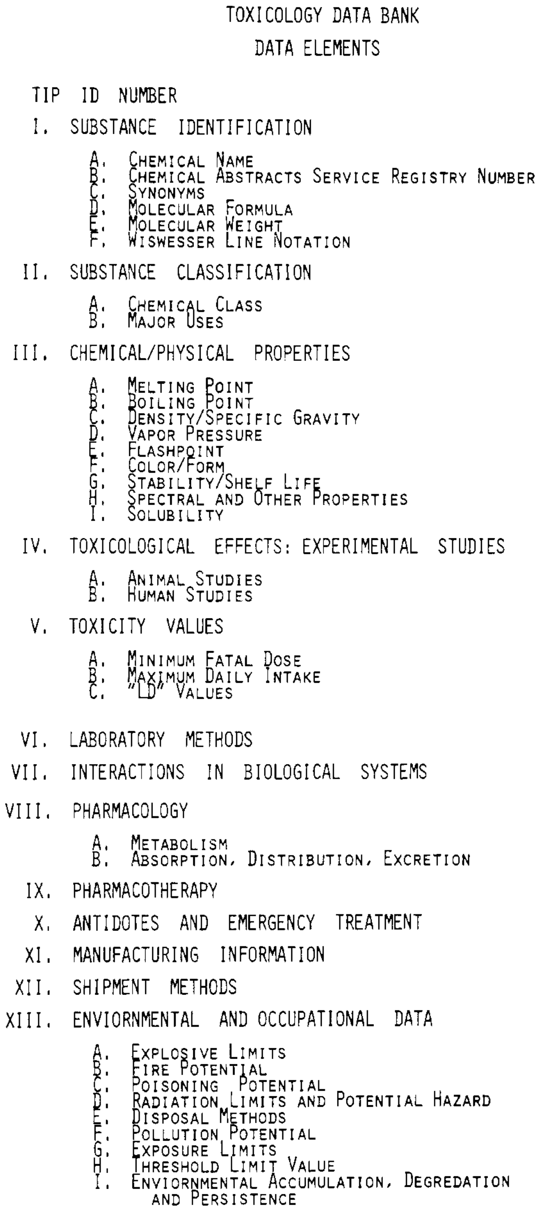
Although the TDB will be used in the obvious manner to retrieve specific data on selected chemicals, other search strategies will be possible. Individual or groups of chemicals could be identified on the basis of user selected criteria. For example, one might request all chemicals that show liver toxicity in humans, have been tested chronically in mice, and are used in the manufacture of plastics. Another capability that should be extremely useful is a form of substructure searching. This will be possible by selecting appropriate fragment codes in the Wiswesser Line Notation field. In addition, such searches could be made based on use class, chemical class, or molecular formula.

In building the TDB, a special activity has been developed for selection of chemicals on which records are generated. Two primary, but not necessarily equally weighted, criteria are used. A chemical can be involved in some reasonable level of exposure to general populations or to specific populations such as an industry or geographical region, and it can have either proven toxicity or be suspect of causing some deleterious biological effect. Because a great deal of impetus for building the TDB has come from other government agencies, many chemicals are selected because of their interests.

The first list of chemicals assigned for data extraction

[†] Presented at the Division of Chemical Literature Program, 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 6–11, 1975.

*Author to whom inquiries should be addressed.



TOXICOLOGY DATA BANK

DATA ELEMENTS

TIP ID NUMBER

I. SUBSTANCE IDENTIFICATION

A. CHEMICAL NAME
B. CHEMICAL ABSTRACTS SERVICE REGISTRY NUMBER
C. SYNONYMS
D. MOLECULAR FORMULA
E. MOLECULAR WEIGHT
F. WISWESSER LINE NOTATION

II. SUBSTANCE CLASSIFICATION

A. CHEMICAL CLASS
B. MAJOR USES

III. CHEMICAL/PHYSICAL PROPERTIES

A. MELTING POINT
B. BOILING POINT
C. DENSITY/SPECIFIC GRAVITY
D. VAPOR PRESSURE
E. FLASHPOINT
F. COLOR/FORM
G. STABILITY/SHELF LIFE
H. SPECTRAL AND OTHER PROPERTIES
I. SOLUBILITY

IV. TOXICOLOGICAL EFFECTS: EXPERIMENTAL STUDIES

A. ANIMAL STUDIES
B. HUMAN STUDIES

V. TOXICITY VALUES

A. MINIMUM FATAL DOSE
B. MAXIMUM DAILY INTAKE
C. "LD" VALUES

VI. LABORATORY METHODS

VII. INTERACTIONS IN BIOLOGICAL SYSTEMS

VIII. PHARMACOLOGY

A. METABOLISM
B. ABSORPTION, DISTRIBUTION, EXCRETION

IX. PHARMACOTHERAPY

X. ANTIDOTES AND EMERGENCY TREATMENT

XI. MANUFACTURING INFORMATION

XII. SHIPMENT METHODS

XIII. ENVIORNMENTAL AND OCCUPATIONAL DATA

A. EXPLOSIVE LIMITS
B. FIRE POTENTIAL
C. POISONING POTENTIAL
D. RADIATION LIMITS AND POTENTIAL HAZARD
E. DISPOSAL METHODS
F. POLLUTION POTENTIAL
G. EXPOSURE LIMITS
H. THRESHOLD LIMIT VALUE
I. ENVIORNMENTAL ACCUMULATION, DEGREDATION AND PERSISTENCE

Figure 1.

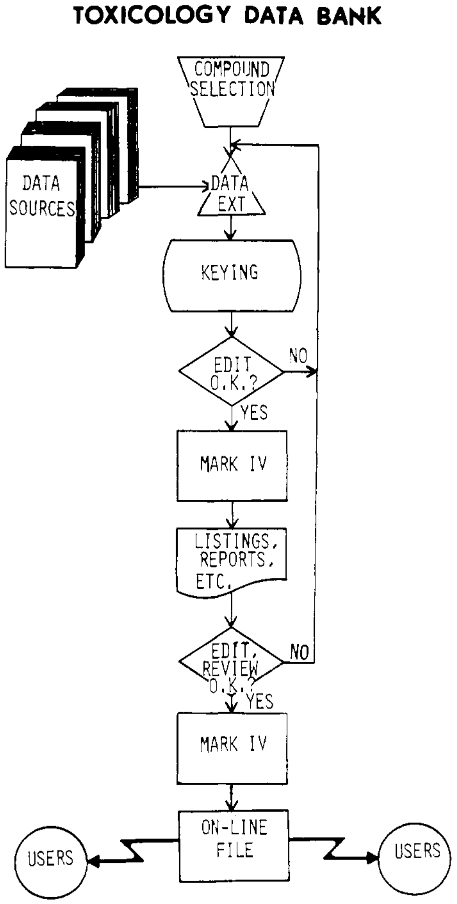
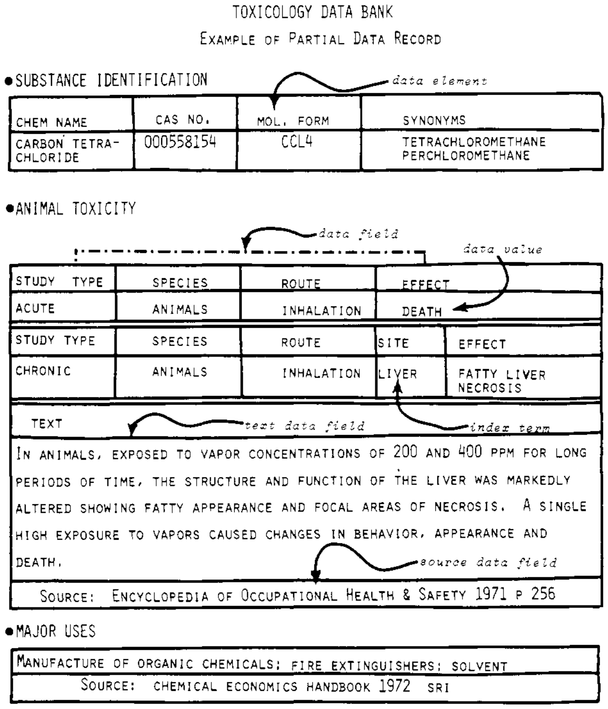


Figure 2.



TOXICOLOGY DATA BANK

EXAMPLE OF PARTIAL DATA RECORD

●SUBSTANCE IDENTIFICATION

CHEM NAME	CAS NO.	MOL. FORM	SYNONYMS
CARBON TETRA-CHLORIDE	000558154	CCL4	TETRACHLOROMETHANE PERCHLOROMETHANE

●ANIMAL TOXICITY

STUDY TYPE	SPECIES	ROUTE	EFFECT
ACUTE	ANIMALS	INHALATION	DEATH

STUDY TYPE	SPECIES	ROUTE	SITE	EFFECT
CHRONIC	ANIMALS	INHALATION	LIVER	FATTY LIVER NECROSIS

TEXT

IN ANIMALS, EXPOSED TO VAPOR CONCENTRATIONS OF 200 AND 400 PPM FOR LONG PERIODS OF TIME, THE STRUCTURE AND FUNCTION OF THE LIVER WAS MARKEDLY ALTERED SHOWING FATTY APPEARANCE AND FOCAL AREAS OF NECROSIS. A SINGLE HIGH EXPOSURE TO VAPORS CAUSED CHANGES IN BEHAVIOR, APPEARANCE AND DEATH.

SOURCE: ENCYCLOPEDIA OF OCCUPATIONAL HEALTH & SAFETY 1971 P 256

●MAJOR USES

MANUFACTURE OF ORGANIC CHEMICALS; FIRE EXTINGUISHERS; SOLVENT

SOURCE: CHEMICAL ECONOMICS HANDBOOK 1972 SRI

Figure 3.

consisted primarily of approximately 250 solvents and monomers. The list now has been expanded to include over 1250 substances, and it is essentially a compilation of lists prepared by other government agencies. Included, for example, are the Chemical Hazards Information Response System (CHRIS) list published by the United States Coast Guard, chemicals included in the Environmental Protection Agency's publication, "Designation and Determination of Removability of Hazardous Substances from Water", and chemicals under consideration by the National Cancer Institute as potential carcinogens.

The cooperative interaction that has resulted is being used to provide assistance in data gathering to the appropriate agencies and to inform them of common interests. Because many compounds are on more than one list, a compilation is being developed that will associate with each selected chemical the agencies which are concerned with it.

To eliminate redundancies, the individual lists were combined into a master compilation of approximately 2800 names. After an initial alphabetic sort and edit, approximately 1700 names remained. This list was processed through the NAME-MATCH program at the Toxicology Information Response Center, Oak Ridge National Laboratory, to assign Chemical Abstract Service (CAS) registry numbers and synonyms to each name. This step generated a list of over 18,000 members with CAS numbers. It was sorted on CAS number, and one name was selected from each group of synonyms to generate the current TDB list. With CAS numbers assigned to each member of the original lists, a matrix of chemical substance and agency interest is being generated for publication. Those names for which no CAS number or multiple CAS numbers were found are being resolved separately.

Although the on-line system has not been implemented as yet, TDB data gathering activities are being used to assist in the preparation of dossiers for the National Cancer Institute (NCI) on substances suspected of being potential carcinogens. These dossiers are prepared by an NCI contractor using TDB records as one source of data. In cases where no record exists, the chemical is assigned for immediate data extraction.

While the primary goal of the program is to develop a data bank that will be responsive to national needs and interests, its worldwide importance also is recognized. A number of other nations and international organizations are either at various stages of developing data banks similar to or complementary to the TDB, or are interested in seeing that one is developed.

An experiment is underway to collect data on selected chemicals through an international collaborative effort. It is being carried out under the auspices of the Organisation for Economic Co-Operation and Development (OECD) in Paris, with the OECD and some of its member nations taking part. Each participating group is selecting approximately 50 chemicals with emphasis placed upon those which are produced in large amounts.

The participants are suggesting data sources, with final selection and assignments coordinated centrally to avoid duplication. Assignment of data elements for extraction is being done on the basis of data sources and specialized interests. All data will be encoded in English and sent to NLM for conversion to machine-readable form, if necessary, and building the chemical records.

One major objective of this project is to determine whether the cooperative establishment of such data bases is more advantageous than separate action at national levels. Moreover, it is hoped that the resulting data will provide some specific results and advantages. For example, the collection could present an international means for dealing with industrial health and safety problems. Of special importance is the fact that it would be of service to all countries rather than select portions being available on a national basis only.

Semiconductor Journals

DONALD T. HAWKINS

Bell Telephone Laboratories, Murray Hill, New Jersey 07974

Received October 20, 1975

Using an on-line literature searching system, the number of papers in many journals dealing with semiconductors was determined. The journals are ranked by the percentage of their contents devoted to semiconductors, and by the total number of semiconductor papers they published. Only four journals devote over half of their contents to semiconductor papers. Approximately half of the 19,646 papers (which were found in 91 journals) appeared in eight journals.

From its beginnings in the late 1940's, semiconductor research and development has grown rapidly; it is now a multimillion dollar effort. Coincident with this growth, semiconductor literature has also burgeoned and diversified. This study represents an attempt to characterize the literature on semiconductors by determining the leading publications in the field. Such information is useful to technical librarians who are faced with rising journal subscription costs and increasingly straitened budgets, as well as to those working in the field who desire to keep abreast of new developments. A list of leading publications is helpful in providing efficient access to much of the new technology. It is also of interest to those who follow the history of science and its network of communication.

The Lockheed Corporation's DIALOG™ system[1] was used to gather the data for this study. Briefly, the DIALOG™ system is an on-line interactive literature searching and information retrieval system which is accessed remotely using a computer terminal and standard telephone line. Currently DIALOG™ supports 16 major data bases. The searcher forms "sets" by entering terms of interest, using a simple command language. Each set formed is given a reference number, and the number of items it contains is reported to the user. The sets can be combined by the Boolean logic operators AND, OR, and NOT. The contents of any set can be displayed immediately at the terminal in a variety of formats, or they can be printed off-line on a high-speed printer. Searchable fields include the title, author(s), keywords, and source for each reference.

For this study, the *Science Abstracts A* (SAA) and *Science Abstracts B* (SAB) data bases were the first choices as indexes to the semiconductor literature. SAA covers the physics and materials science aspects of semiconductor materials, while SAB covers the electrical engineering and device aspects. Together, they would be expected to provide thorough coverage of the world's literature on semiconductors. At the time of this study, the data for 1970 through April 1975 were available on the DIALOG™ system.

By the time this study had been completed, a total of 91 physics or materials science journals publishing acticles on semiconductors had been identified. Many of these were chosen from previous journal network or journal hierarchy studies.[2–4] Some were selected after consultation with experts in the field, and a few became evident as this study proceeded. The list of journals arranged by broad subject coverage, which appears in the yearly guide to the *Science Citation Index*, was also checked.[5] Attempts were made to ensure that no major journals were omitted from this survey, by scanning several printed issues of SAA and SAB.

The search procedure with the DIALOG™ system was as follows. A set was formed containing items with terms of the form SEMICOND... in either the SAA or SAB title or keyword (the DIALOG™ system's descriptor and identifier) fields. Sets were then formed containing all papers published by each of the 91 journals. A special attempt was made to include all of the many variations of journal title abbreviations, etc. (The ASTM CODEN was used where possible, but it has been included in SAA and SAB only since 1973.) Titles for Russian journals which are translated cover to cover were included in both the original and translated forms. (All representations of a title were ORed together into one set, eliminating duplicates.)

Each journal title was then combined in turn in a Boolean