ACKNOWLEDGMENT

This work was performed for the Electric Power Research Institute under Contract No. RP 1643-1. The Project Officer was Dr. J. McCarroll.

REFERENCES AND NOTES

(1) Hansch, C.; Leo, A. J. "Substituent Constants for Correlation Analysis in Chemistry and Biology"; Wiley: New York, 1979.
(2) Hansch, C.; Clayton, J. M. "Lipophilic Character and Biological Activity of Drugs II: The Parabolic Case", *J. Pharm. Sci.* **1973**, *62*, 1-21.
(3) Hansch, C.; Dunn, W. J. "Linear Relationships Between Lipophilic Character and Biological Activity of Drugs", *J. Pharm. Sci.* **1972**, *61*, 1-19.
(4) Higuichi, T.; Davis, S. S. "Thermodynamic Analysis of Structure-Activity Relationships of Drugs: Prediction of Optimal Structure", *J. Pharm. Sci.* **1970**, *59*, 1376-1383.
(5) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. "Computer Assisted Studies of Chemical Structure and Biological Function"; Wiley: New York, 1979.

EPA Health and Environmental Effects Data Analysis System†

DAVID LEFKOVITZ*
University of Pennsylvania, Philadelphia, Pennsylvania 19104

AMY RISPIN
U.S. Environmental Protection Agency, Office of Pesticides and Toxic Substances, Washington, DC 20460

CAROL KULP and HELEN HILL
University of Pennsylvania, Philadelphia, Pennsylvania 19104

Received October 9, 1980

This paper discusses the development of a system to organize, store, retrieve, and correlate data pertaining to chemicals and their biological and environmental effects. The particular problems of data identification, acquisition, classification, and automation are discussed in relation to existing data sources and methods of data collection and analysis. The problems of computer software development are also addressed, and a design overview of the system is presented.

INTRODUCTION

The Office of Toxic Substances (OTS) of the U.S. Environmental Protection Agency (EPA) is charged with making regulatory decisions under the Toxic Substances Control Act (TSCA) concerning the 43 000 commercial chemicals listed in the TSCA Inventory.[1] Under various sections of the TSCA, chemicals in the inventory must be ranked for regulatory concern or selected for testing. In the course of assessing the toxicological hazard of chemicals for regulatory purposes, the computer can be used as a tool by the skilled scientist. For this reason, the Office of Pesticides and Toxic Substances (OPTS) is developing the Health and Environmental Effects Data Analysis (HEEDA) system. In providing for structure–activity prediction in toxicology, the HEEDA system contains validated or reviewed toxicological data that can be correlated statistically with structural features of chemicals. The regulatory scientists in the OPTS plan to use the techniques of quantitative structure–activity relationships (QSARs) to focus attention within the agency on chemicals of concern. Since QSARs in toxicology represent a science in its infancy, the HEEDA system will be used to assess the limitations as well as the scope of QSAR prediction.

In the creation of a software system for the processing of biological and chemical structural information, the initial phase of development must focus on data acquisition and organization (Figure 1). In this initial phase, appropriate sources of validated data must be identified for inclusion in the data base. The data must be classified biologically and chemically for ease of organization and retrieval by the analyst.

† Presented on April 23, 1980, as a part of the Symposium on Devlopment and Use of Reliable Data Bases for Quantitative Structure–Activity Relationships (QSARs) during the 14th Middle Atlantic Regional Meeting of the American Chemical Society, King of Prussia, PA.

In the data access phase of development, the system is designed and programmed to provide a vehicle for information retrieval. The ability to perform substructure searches is an essential feature of data organization for retrieval of information in the chemically oriented data base. As a node in the Chemical Substances Information Network (CSIN),[2] the HEEDA system will acquire its substructure search capabilities from the CSIN Chemical Structure and Nomenclature System (CSNS).[3]

The final phase of system development is one of data correlation methodologies. HEEDA will employ two mechanisms for data correlation. The first is that of report generation to provide graphical and visual display of information in a variety of formats. The second method of data correlation is by means of mathematical modeling techniques from the discipline of QSARs.

The development of the HEEDA system has been directed toward the creation of a computer environment that contains the necessary components for structure–activity experimentation and prediction in areas of regulatory concern. At the heart of the HEEDA system is a collection of standardized, reviewed data that can be subjected to various statistical methods to correlate biological end effects with structural features. From these biological data, sets of chemicals can be assembled that are well characterized with respect to the biological effect of concern. The Office of Toxic Substances of EPA is organizing and validating many such data sets. The potential hazard of uncharacterized compounds will be assessed by comparison with the data in the training sets. Authenticated training sets can be used with different correlative techniques to test the validity of the statistical models. In the context of reliable data for training sets, different chemical–structural descriptors can be tested for their usefulness in structure–activity prediction.

0095-2338/81/1621-0018$01.25/0

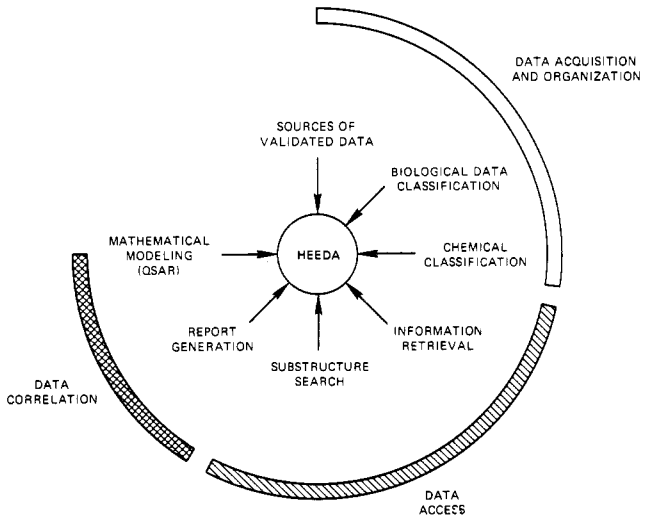

Figure 1. Diagram of the major phases of development of the Health and Environmental Effects Data Analysis (HEEDA) system.

Key considerations in the assembly and validation of toxicological and physical–chemical data for the chemicals represented in HEEDA are the details of experimental protocol as to completeness, statistical validity, and conformity to current laboratory practices. The review and validation of toxicological data are resource intensive. As these data are assembled in the integrated data base of the HEEDA system, necessary additional areas of data collection or review can be identified.

SCIENTIFIC APPROACH TO THE SOLUTION

Sources of Data. Initially the focus of our structure–activity research was in the area of prediction of carcinogenicity from chemical-structural features. It became apparent that a reliable carcinogenesis data base was needed to act as a training set. When we examined the chemicals reviewed in the International Agency for Research on Cancer (IARC) Monographs[4,5] and tested in the 2-year rodent bioassays of the National Cancer Institute (NCI),[6,7] it became clear that these sources of data alone could not provide us with a training set that was balanced with respect to (1) positive and negative end effects, (2) different structural classes, and (3) animal vs. human data. Therefore the cancer file will be extended to include the chronic toxicity tests on *N*-nitroso compounds of Lijinsky.[8] It will also contain those polyaromatic hydrocarbons and aromatic amines from the scientific literature that have been extensively characterized as carcinogens in the laboratory.

It is desirable to be able to augment the cancer file with antineoplasticity and mutagenicity data. The NCI Division of Cancer Treatment antitumor data base[9] developed under the supervision of S. Richman is being added to HEEDA. HEEDA will also contain the results of the 24 GENE-TOX[10] mutagenicity review panels sponsored by EPA. In addition to the ability to model structural correlates with each of these end effect types, the HEEDA system provides for correlation of one biological effect with another. For example, mutagenic activity can be compared with carcinogenicity.

Another major area for structure–activity correlation is toxicity. The ecotoxicological screening data from the U.S. Fish and Wildlife Service[11,12] are being entered into HEEDA. These files include mammalian, avian, and aquatic toxicity testing. A U.S. Fish and Wildlife Service house evaluation code will indicate quality and completeness of protocol for each chemical entered in this file. The HEEDA data base will include other toxicity files as well. A neurotoxicity file being prepared at EPA will be included. Metabolite data have been identified to augment the information about chemicals in the ecotoxicology and the *N*-nitroso files. In addition, an environmental fate file is being developed for EPA at Syracuse Research Corporation. This file includes physical chemistry data as well as results of field studies in biodegradation.

The utility of the octanol–water partition coefficient in providing correlation of chemical structure with toxicity has been demonstrated. Therefore, the partition coefficients from the Pomona Med-Chem project[13] are being added to HEEDA. In addition, HEEDA will contain validated computer routines for computation of partition coefficients.

System Organization for Chemical and Biological Information. The HEEDA data base design is unique in that it combines in one system, in an open-ended manner, meaningful observational categories of information of regulatory concern. The data base organization is illustrated in Figure 2. All information in the system is related to a specific agent that is uniquely identified by its CAS registry number. This number points to two file systems. One is called the Auxiliary Chemical Data File; the other is called the Experimental Data File. The former contains chemical data that are required for identification (names), modeling (classes and substructures), and cross-referencing to other related chemicals, such as metabolic or reaction products.

The second file system contains all of the experimentally derived data, organized hierarchically, as shown in Figure 2. At the top of the hierarchy is the unique chemical identifier, the CAS registry number. At the second level are distinguished a series of observational classes. Four such classes are shown in the figure: end effects, environmental measurements, biological measurements, and chemical/physical properties. More classes can be added to the data base as needed. The shadow boxes in the figure indicate that, at any level, any number of repeated items of data can be entered. Thus, at the top level, the total data base is characterized as a series of CAS numbers. At the second level a series of observational classes can be entered for each CAS registry number (as noted above). The third level divides each observational class into subclasses. The figure shows that the observational class *end effects* is divided into *mutagenesis*, *carcinogenicity*, *toxicity*, etc. As another example, *environmental measurements* are subdivided at the third level, and subdivisions of the remaining observational classes are represented as dashed lines. The fourth and fifth levels represent a single reference (paper) or documented test. Thus, any number of tests can be recorded per observational subclass. The fourth level presents only the summary results of the test, and the fifth level presents the remaining experimental and other descriptive details through a series of data items called qualifiers. As will be explained below, the qualifiers are organized along the lines of a scientific paper, presenting the citation of the work, methods and materials, detailed results, and, optionally, discussion.

This organization enables searches of the following kind to be performed:

(1) find summary results for all observational data pertaining to chemical X,

(2) find summary results for all end effects of chemical X,

(3) find all carcinogenic summary results and qualifier details for chemical X,

(4) find all carcinogenic summary results on the female rat species for chemical X.

The HEEDA system is designed to accept data with varying degrees of detail and format at the qualifier level, depending upon the scientific needs of the source. In short, qualifier definitions are open ended and can be expected to run into the thousands. Therefore some data classification is necessary at the qualifier level so that the scientist can deal effectively with

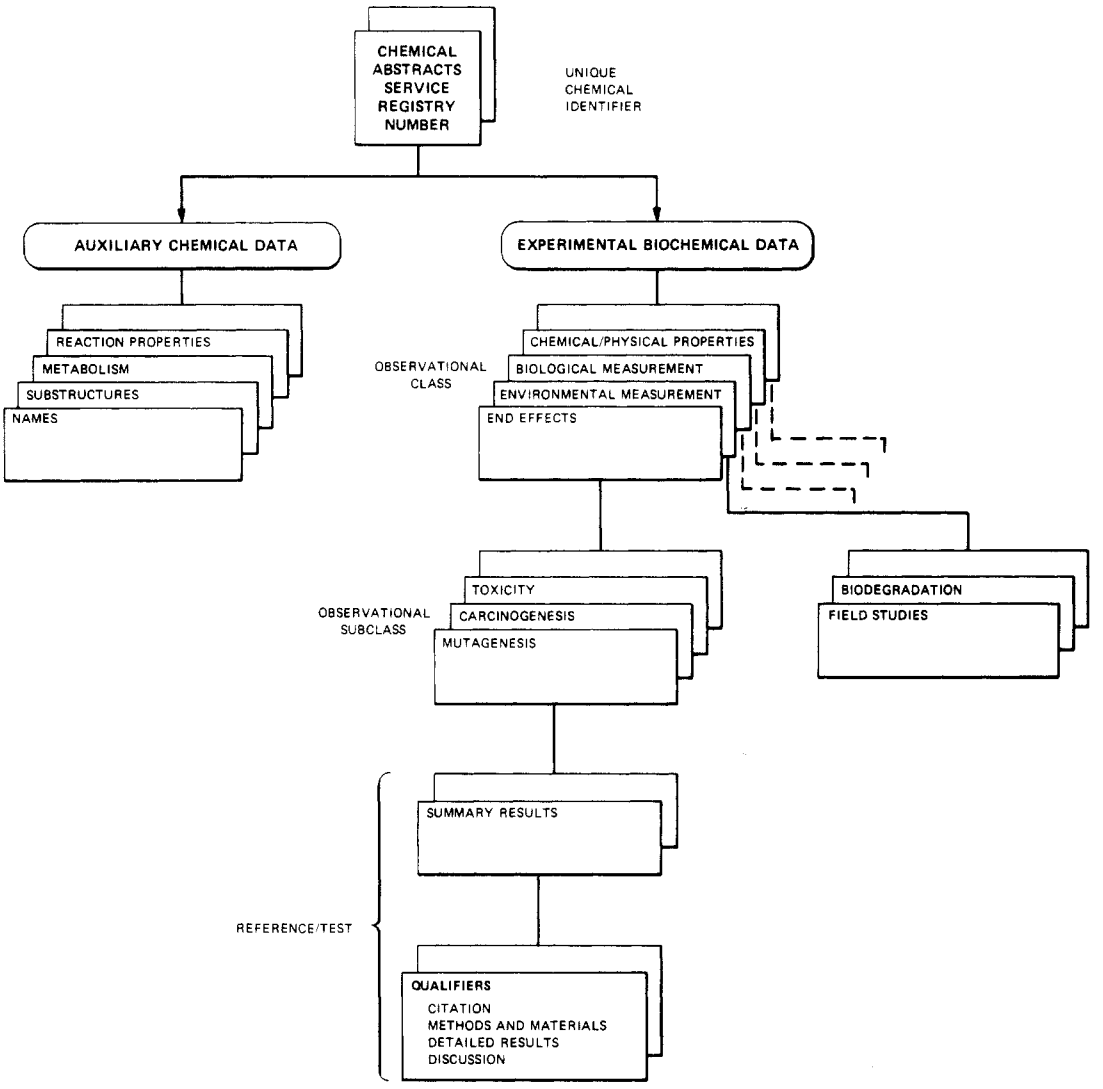

Figure 2. Organization of the Health and Environmental Effects Data Analysis (HEEDA) data base.

the data base for the purposes of file search, organization of displays and reports, and extraction of quantitative data to be used in calculations. However, deep classification hierarchies have the disadvantage that information becomes highly differentiated and cross-referencing becomes difficult. A method was sought that would create a shallow but familiar and conceptually simplified classification scheme in which it would be relatively easy to tie relevant data together at the lower levels of the hierarchy. This method would enable the user to readily identify analogous types of data elements across all test systems, even if these data elements differ widely in specific details. For example, the user may wish to retrieve all compounds with both carcinogenic and mutagenic tests performed in the same organisms. He would like to display all organism data, the test agent, metabolic precursors and products, controls, dose administration data, and results and references.

The classification method developed to answer these needs spans the summary results and qualifier levels. It parallels the general outline of a scientific paper:

citation	
test system and summary results (abstract)	elements in HEEDA data base
detailed results	
discussion	optional
bibliography	not included

The HEEDA files will store discussion as comments, when appropriate, but will not store the bibliography, since references can readily be obtained from source documents if necessary. As described, the files are so organized that all information pertaining to a particular test is presented at two levels of specificity. The first and higher level contains the test system and summary results in a standard format. The test system is selected from a controlled vocabulary. Examples are

mutagenicity—effects on chromosomes
mutagenicity—gene mutations
toxicity—lethality
carcinogenicity—long-term bioassay

The summary results are represented by controlled formats, such as

+	positive
–	negative
+?	questionable positive
V	variable or borderline results
I	insufficient evidence for conclusions
LD_{50} 5 mg/kg	(self-explanatory)

The more specific level contains an open-ended set of data items called qualifiers. Onto this level are mapped the citation, methods and materials, detailed results, and discussion (or comments), as shown in Table I. The qualifier data are classified by a three-level hierarchy, each level of which is called a *facet*. The first facet corresponds to the major categories of the scientific paper. The second facet defines a

Table I. Three Facets of HEEDA[a] Data Base Organization and Examples of Use

section of scientific paper	level		facet		
	no.	name	1	2	3
abstract	4	summary results	*b*	*b*	*b*
citation	5	qualifier	reference	primary source	author
				standard source	year
					standard citation
					other citation
					laboratory manual
					EMIC[c] number
materials and methods	5	qualifier	substance	test agent control	name
					concentration
				selective agent	lot number
					purity
			organism	test agent	common name
				control host	scientific name
					source
					numbers
					age
			conditions	administration	route
					duration
				environmental	type
			methodology	statistical	method name
					method description
				analytical	method name
					method description
detailed results	5	qualifier	quantitative results	test agent control	number of mutants per plate
					LD_{50}
					mean
			qualitative results		standard deviation
discussion	5	qualifier	test comment	*b*	*b*
				b	*b*

[a] Health and Environmental Effects Data Analysis system. [b] Not applicable. [c] Environmental Mutagen Information Center, Oak Ridge National Laboratory.

subcategory or role within the first facet. For example, as shown in Table I, a methodology can be statistical or analytical; the role of an organism can be test agent, control, or host. The third facet contains the actual qualifier data. Table I presents examples of the values that these facets can assume. The number of facets is fixed, but the values within the facets are open ended in order to allow retention of any desired degree of experimental detail.

For inquiry the user can specify any combination of the three facets, and all data items so referenced will be retrieved. For example, if only the test agent value of the second facet were cited, all facet 1 and facet 3 qualifier names with "test agent" in the second facet would be accessed.

In summary, for HEEDA to serve the needs of regulatory and research scientists it must have certain open-ended features in order to accept data from a variety of sources. In these sources, the recording of experimental information is not necessarily subject to control either by HEEDA management or by HEEDA users; where such control may exist, the source file designers must still have considerable latitude in determining the types and detail of data to be stored. Conversely, for HEEDA to be a useful scientific tool for data retrieval and structure–activity correlation, it must be able to identify common or comparable elements of data, where they exist, across all sources. These open-ended features are (1) ability to define any number of observational classes and subclasses per class, (2) ability to store any number of test references per subclass, and (3) ability to define any number and type of qualifier details per test reference.

Chemical Classification. The ability to classify chemicals by substructural features and to compare biological effects within and between such classes is an important aspect of the assessment of chemical hazard for purposes of testing and regulation. In addition, in order to select appropriate training sets for use in structure–activity relationships modeling, the ability to select and sort by chemical structural features is imperative (see section on modeling). HEEDA provides for two types of such chemical classification: *a priori* and *ad hoc*. A priori classifications are those that have been developed by an intellectual process to express some scientific observation. The Department of Health, Education, and Welfare developed a scheme of chemical classes that was augmented for use in the GENE-TOX program.[14] This classification currently has 69 classes, such as "acyl halides", "aldehydes", "aromatic azo compounds", and "quinones". This classification scheme is being entered into HEEDA with the GENE-TOX chemicals. The class names are coded and entered into the Names record illustrated in Figure 2. That is, they are considered to be a "name", except that they refer to a class instead of a unique compound. In addition, they are identified by their source (in this case GENE-TOX). In this way any number of class names can be assigned per compound and separately searched or retrieved.

Ad hoc classifications are those in which a computer assigns substructural fragments to a chemical, generally from a connection table. Classification may be made either algorithmically (by following a set of rules) or by matching the structure to a predefined dictionary of substructures. These class identifiers or "keys", as they are called, are stored in the substructures record of Figure 2. Again, any number of different algorithmic assignment systems or dictionaries can be used. At present, the HEEDA system contains a set of programs that assign TSS keys.[15] Other potential candidates are BASIC,[16] CIDS,[17] NCI,[18] and WLN[19] keys. These are used by the mathematical modeling programs as structural features.

Data Definition. Once data have been classified as described above, there remains the task of representing the actual data in the computer system. For this purpose a number of accomodations are required, because people can accept far less

formal data format description than can the computer. A three-step procedure has been developed for this purpose.

In step 1 the data element and its source are identified and given a natural language description. It is then assigned a faceted code.

In step 2 the data element is assigned a set of attributes:

Attribute	Explanation
Search/print	Searchable data values must be standardized. Print-only data need not be standardized.
Mode	Data may be alphanumeric, decimal, or scientific notation.
Range	Numeric data sometimes are expressed as ranges (e.g., "1–3 days").
Table	Data of any mode may be entered as a "Table", or string of values separated by commas (e.g., a series of concentrations could be "1, 5, 10, 25 μg/mL").
Subscript	Elements may be repeated (e.g., an experiment may be repeated at three different temperatres; thus the code for temperature is subscripted, and the subscript values go from 1 to 3).
Units	Certain numeric data require units.
Decode table	Data can be encoded to reduce storage and search time. These are then decoded by a table for display.
Validation table	Some data are checked at time of input for a predetermined set of allowable values. These are stored in the Validation table.

In step 3 the code along with its source, description, and attributes is entered into a dictionary called the Data Definition Table. When the code is the same as an existing one, only the new source need be added. Because an objective of HEEDA is to create common data interpretation across sources and references, this may happen frequently.

The Data Base Manager is responsible for the control of the hierarchic classification and the data element dictionary (see Data Base Management and Dictionary Control).

TECHNICAL APPROACH TO THE SOLUTION

The HEEDA System. A system that will implement the solution described above must have the following general characteristics: (1) It must accept data from a large number of sources in various formats, while permitting coherent search across these sources. (2) It must enable data to be efficiently retrieved according to the biological and chemical classifications described above. (3) It must provide the necessary search and correlative tools to carry out the principal scientific functions of the system.

The most important system design features that are needed to implement the solution described above are the record, data base, data access, and data correlation designs.

Data fields are self-defining within the record rather than having predetermined positions and formats. This means that any HEEDA record can be updated with fields from existing or new data sources without reformatting the file, and a completely open-ended set of data items can be accommodated. All fields have variable length data expressed in simple numeric, numeric range, tabular, or textual format. Any numeric field can have units, and unit translations are provided in a table to enable a multiplicity of units to be specifiable for retrieval or display. A data base administrator has the responsibility to assign common data field names across sources, wherever possible, in order to enhance the coherency of data access throughout the system. The HEEDA data base design follows the hierarchic structure illustrated in Figure 2.

Data can be accessed from the HEEDA data base in three modes: (1) CAS registry number, (2) predefined chemical class names, and (3) a Boolean combination of HEEDA data fields, in which any field can be required to be "equal to", "less than", "greater than", etc., a particular value. In each mode any data field can be selected for display or for forwarding to a report generator. Specially formatted quantitative values in the summary results level can be forwarded to a model. As discussed above, data must also be accessible by substructure for creation of ad hoc structural classes. To achieve this, HEEDA will be interfaced to the Chemical Structure and Nomenclature System (CSNS) via the Chemical Substance Information Network (CSIN) in order to provide substructure and nomenclature search.[2,3]

To perform data correlation, an analyst interacts with the system in two steps. First, by one of the data access methods he selects data for review or correlation. Second, he selects an appropriate method of correlation. Two general correlation tools are provided, Reports and Models. Reports provide a range of correlation techniques based primarily on visual effect. Data can be sorted, columnated, arrayed into matrices, and graphed. Models are constructed on the basis of mathematical or statistical procedures that operate on continuous numeric or discrete data. In general, end effects are viewed as dependent variables, while physical–chemical and structural descriptors are viewed as independent variables.

Figure 3 presents a block diagram of the HEEDA system.

(*1*) *Input Files.* Data can be made available to HEEDA in two ways: by preexisting automated format (represented by a tape symbol), which must be converted into the storage format of the HEEDA data base, or by newly developed source data than can be entered according to a HEEDA-specified format (card symbol).

(*2*) *File Maintenance Module.* The File Maintenance module performs three functions: conversion of input files into the HEEDA data base [from (1)], modification of records already in the data base [from (3)], and incorporation of chemical descriptors generated automatically from special purpose programs [from (4)].

(*3*) *On-Line File Maintenance.* This program enables a person to examine HEEDA records from the data base by CAS registry number and update them by addition, modification, and deletion of individual data fields or entire tests. This function is, of course, tightly controlled.

(*4*) *Automatic Descriptor Assignment* (*ADA*). There are a number of programs available that can generate structural or chemical descriptors from a connection table representation of the molecule. HEEDA will provide the CAS Registry III Connection Table to these programs. The outputs of these programs would be formatted by the File Maintenance Module in such a way that the model programs could select any type of descriptor that would be suitable for the particular model.

(*5*) *HEEDA Data Base.* The HEEDA data base has been described and illustrated in Figure 2. The system at present provides random access only by means of a concatenated key:

CAS registry number
Observational class—subclass
Test record number

The user can select any level of this key and obtain a selection of all data heirarchically below the stated key level. Thus, all data pertaining to a specific agent down to the data pertaining to a specific test within an observational class can be obtained interactively. Generic searches of the file, based upon logical combinations of values of observational class,

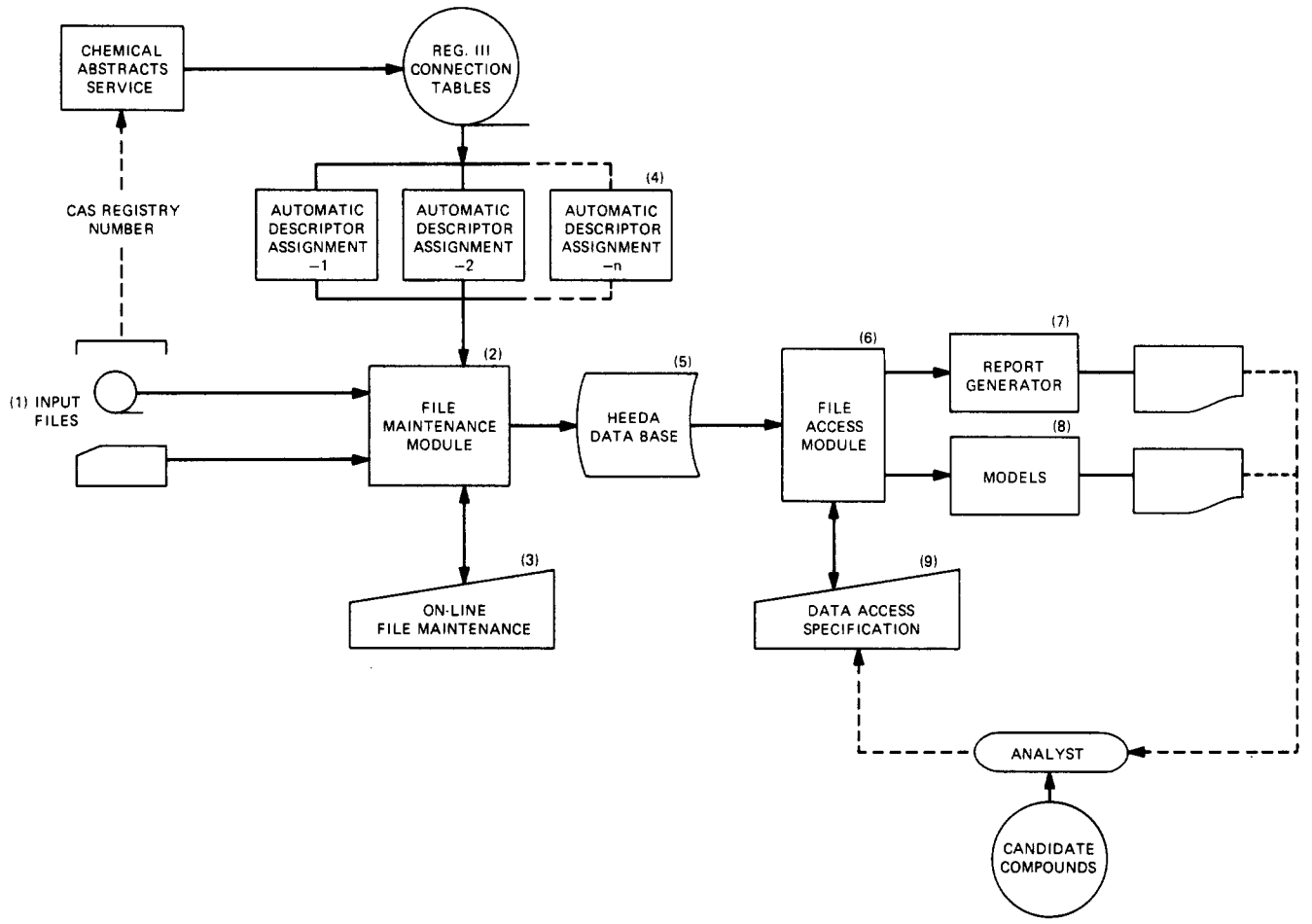

Figure 3. Block diagram of the Health and Environmental Effects Data Analysis (HEEDA) system. The automatic descriptor assignment includes such elements as substructures, partition coefficients, and steric descriptors.

subclass, test name and results, and qualifiers, are obtained by a serial search of the file.

(*6*) *File Access Module.* The File Access Module enables the analyst to select subfiles based upon any combination of qualifiers and end effects for input to the report generator and models. The system will retrieve records based upon a Boolean logic expression. The variables of the Boolean logic used for retrieval can have any of the following forms:

(i) A particular type of data field is present, such as "species".

(ii) The data in any type field equals a given value (for example, "species" = "rat").

(iii) The data in a numeric field satisfies a numeric comparison (=, <, >, ≠, etc.).

(*7*) *Report Generator.* The Report Generator enables the analyst to correlate data extracted by the File Access Module, by sorting and visual means. Three types of report organization are provided: columnar, matrix, and graphical.

(*8*) *Models.* Models constitute the other major output function of HEEDA. Mathematical models establish some functional relationship between readily available descriptive properties of a chemical and its biological activity. The estimate or prediction of activity of a candidate or unknown chemical then involves selection of the appropriate model based upon the candidate's chemical class and the activity to be estimated.

HEEDA will also facilitate the development and use of manually produced classifications or decision trees. In such modeling, chemicals are divided, usually by an intellectual process, into a hierarchy of classes and subclasses on the basis of what the classifier believes will define groups of compounds having like activity. Many such trees can thus be developed for various types of biological activity on the basis of the requirements of a particular investigation and available data. HEEDA will provide the means to store these trees and to automatically place any candidate compound in the appropriate groups of selected trees. The groups (i.e., the classes and subclasses) are defined, as in the case of mathematical models, by sets of structural and physical descriptors; residence of a chemical in a particular group implies activity similar to other members of the group.

(*9*) *Data Access Specification.* The analyst, shown in Figure 3, will use HEEDA in a two-step manner. First, a specification must be made for data to be accessed from the file and a report format or model specified. This procedure is performed interactively from a terminal. The analyst can also display any selection of data related to a particular agent (by CAS registry number) on the terminal. In the second step, the search of the entire file, based upon the data access specification, and the production of a report or model are performed as batch operations.

Data Base Management and Dictionary Control. HEEDA must maintain data dictionary control at several levels. This is the responsibility of a data base manager. These levels of control are defined as follows:

(1) The same item in different observational classes should be similarly defined.

(2) For each data item a decision must be made as to whether its values must come from a standardized and controlled vocabulary or whether values are freely accepted from their source. The former is necessary to enable a complete search across all sources, levels, and classes of information. The latter will generally suffice if the data are to be used only for display. Sources of current and prospective HEEDA

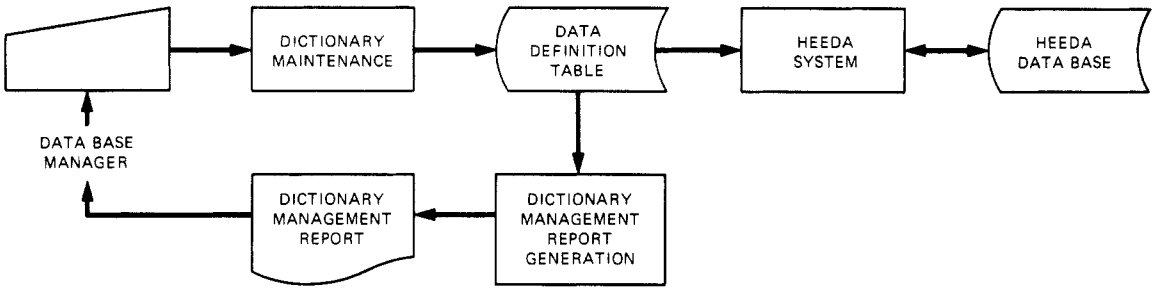

Figure 4. Health and Environmental Effects Data Analysis (HEEDA) system data base management and control via the data definition table.

dictionaries include International Register of Potentially Toxic Chemicals,[20] Environmental Mutagen Information Center,[14] the Toxicology Data Management System of the National Center for Toxicological Research,[21] and the U.S. Fish and Wildlife Service Toxicology Data System.[11]

(3) Each data item must be classified by facets so as to be appropriately grouped with other items.

(4) Each data item must be identified by its sources.

(5) Each data item must be assigned the attributes described in Data Definition.

The controls are stored in a data definition table, as shown in Figure 4. The data base manager maintains this table and can obtain a report on its content. This table is then used directly by the various HEEDA input, search, reporting, and modeling programs for interpreting the data base.

Substructure Search: HEEDA as a Component of CSIN. A necessary function for HEEDA users is the ability to perform substructure research. This enables one to define ad hoc classes via arbitrarily specified substructures. The normal mode of operation would be to submit a substructure search to find the CAS registry numbers of all compounds in a particular data base of interest, such as CANCERLINE, TOXLINE, or the TSCA Inventory, and to extract the desired data for these compounds from HEEDA. It is envisioned that substructure search will be provided to a consolidated set of chemical files of common interest to government agencies and many research and industrial organizations by means of the Chemical Substances Information Network.[2] This system, currently under development, will interface a number of publicly accessible systems and data bases by means of a data communications network and a common front-end inquiry capability. A specific system, the Chemical Structure and Nomenclature System,[3] will act as a central directory of chemical substances that are referenced in any of the CSIN component systems. The CSNS would perform two basic functions. First, it would provide full or substructure search on the total set of CSIN compounds, and, second, it would automatically transmit the responding CAS numbers to any of the CSIN component systems for further processing. In this way, HEEDA, as a CSIN component, would incorporate the functions of full structure identification via nomenclature or structural formula and of substructure search, thus providing the necessary ad hoc classification capability.

Data Correlation. The ultimate objective of HEEDA is to provide a tool that may assist scientists in discovery of why chemical substances behave as they do. The basic mechanism of this tool is correlation of known facts to produce new knowledge. To do this effectively the facts must first be accumulated and stored in a format that is suitable for the correlation process. These features of the HEEDA system have been described above. HEEDA must also provide a range of correlation techniques.

Report Generation. A report enables one to correlate data in several ways. First, data can be selected and juxtaposed across various levels of the data base illustrated in Figure 2. Second, the visual organization of information in the report itself is an effective correlation aid. This is accomplished by sorting to combine and coordinate like information on the page and by the report format.

An example of part of a report within the observational class, end effect, and the subclasses, mutagenesis and carcinogenesis, is shown in Figure 5. The mutagenesis test systems selected were CHO (chinese hamster ovary cell culture), WP2 (*E. coli* WP2 reversion test), HT (heritable translocation in mice), and L5178Y (mouse lymphoma). The first column of the report contains chemical class, the second column the name of the chemical tested, and the third column its CAS registry number. The report is sorted by chemical class, and within each class chemicals are sorted by name. Test results are displayed under "**" for the four mutagenesis test systems and carcinogenesis results, along with activator indicator, AC, for CHO, WP2, and L5178Y (NA = no activator, S9 = S9 mixture, NI = activator, no inducer), germ cell stage tested, STAGE, for HT (1 = spermatogonia, 2 = spermatocyte, 3 = spermatid, 4 = spermatozoa), locus, L, for L5178Y (T = TK +/-), and source, SR, for CAR (T = Tomatis, N = NCI bioassay).

Mathematical Models. One purpose of HEEDA is to serve as a laboratory for testing computer models for quantitative structure–activity relationships. These models assign weights to various physical or structural–chemical descriptors on the basis of a correlation between the occurrence of specific instances of these descriptors and the known value of activity or end effect of the compounds in which they occur. This selection of the characteristic descriptors and the assignment of relative weights is called the model, and the set of compounds with known end effects that are used by the algorithmic process to generate the model is called the training set. The chemical descriptors are the independent variables, and the chemical activity or end effect is the dependent variable. The model is used to estimate the activity for a candidate compound based upon the occurrence and weighting of the descriptors in the new compound.

In addition to the specific correlation algorithm employed in the computer program available for modeling, the success of the modeling process will depend upon certain other factors. First, and most important, are the biochemical definitions and meaning of the independent variables. Second is the development of a modeling algorithm designed to handle the number of independent variables in the problem. Third is the design of an algorithm to handle the two commonly encountered types of variables, continuous and discrete. Examples of the former are molecular weight, partition coefficient, and toxicity, expressed as LD_{50}. Examples of the latter are carcinogenicity, expressed as "+/-/?" or "the existence or nonexistence of substructural fragment X".

There are many computer algorithms for constructing these models. A goal is for HEEDA to provide the three necessary facilities to test and comparatively evaluate both the modeling programs and the descriptor types. These are (1) a body of reliable end effects, (2) a variety of methods for generating and storing different chemical, physical, and structural de-

05/12/80

TEST SYSTEM SUMMARY REPORT

•••••CLASS•••••	••••••••NAME••••••••	••CAS#•••	•CHO• ** AC	•WP2• ** AC	•••HT••• ** STAGE	•L5178Y• * L AC	•CAR• ** SR
ALKALOIDS	14ALPHA,19-DIHYDRO-1 2,13ALPHA-DIHYDROX Y-20-NORCROTALANAN -11,15-DIONE	315220		- NI - NI			+ T
ALKYL EXPOXIDES	ENDO,ENDO-1,2,3,4,10 ,10-HEXACHLORO-6,7 -EPOXY-1,4,4A,5,6, 7,8,8A-OCTAHYDRO-1 ,4:5,8-DIMETHANONA PHTHALENE	72208		- NI			? N
ALKYL SULFATES, SULFOXIDES,SU LFONES,SULFON ATES	DIETHYL SULFATE	64675	+ NA				+ T
	DIMETHYL SULFATE	77781	+ NA				+ T
	ETHYL METHANESULFONA TE	62500	+ NA		+ 43	+ T S9 + T NA	+ T
	ISOPROPYL METHANESUL FONATE	926067	+ NA		+ 43		
	METHYL METHANESULFON ATE	66273	+ NA	+ NA + NA	+ 32	+ T NA + T S9	
	N,N-DIETHYL-5-METHYL (1)BENZOTHIOPYRANO (4,3,2-CD)INDAZOLE 2-ETHANEAMINE MONO METHANESULFONATE	52871235				+ T NA	
	O,O-DIETHYL O-(4-MET HYLSULFINYL)PHENYL THIOPHOSPHATE	115902		- NI			

Figure 5. Sample of a report generated by the Health and Environmental Effects Data Analysis (HEEDA) system.

scriptors, and (3) a means of accumulating and evaluating performance information on the models so that standards against which to measure new methods can be established. The various models may also turn out to have complementary advantages, so that the development of a battery of models to produce a single decision may be desirable.

Table II presents a list of some of the types of programs available to HEEDA. Their essential characteristics in terms of the above discussion are also presented in the table.

FILE CONTENT

Carcinogenesis Bioassay. Two major sets of reviewed carcinogenicity data are being entered into HEEDA, the NCI rodent bioassay results as reviewed by the NCI Chemical Clearinghouse and published in NCI Technical Reports and reviews of existing carcinogenicity literature performed by the International Agency for Research on Cancer (IARC) of the World Health Organization (WHO). In addition, data are being entered from the extensive screening tests in rats of *N*-nitroso compounds from the laboratory of William Lijinsky, Frederick Cancer Research Center (FCRC), and Oak Ridge National Laboratory (ORNL). Many of the Lijinsky *N*-nitroso studies as well as several NCI bioassay results have been reviewed by the IARC panels. Each of the three files is entered in its entirety. Because of the structure of the data base, duplicate studies or reviews of the same chemical are explicit to the user.

(*1*) *NCI Bioassay Program.* About 200 chemicals have been fully tested, reviewed, and published from the NCI bioassay program. These tests were performed in two species of rodent with relatively uniform protocols.[7] The chemicals tested in this program have been selected by the Chemical Selection Working Group of the NCI Chemical Clearinghouse. They are all chemicals in commerce and run the gamut from drugs to pesticides and chemicals of industrial importance. Chemicals tested in the NCI bioassay program were not always purified. Often they were technical grade such as dicophol (40–60% pure) or mixtures such as APC (aspirin, phenacetin, and caffeine).

Griesemer and Cueto[6] have recently published criteria for very strong, sufficient, and equivocal evidence of carcinogenicity and assigned the chemicals tested in the bioassay program to nine categories in decreasing order of evidence for carcinogenicity.

Results for these NCI bioassays have been entered into the HEEDA data base from Griesemer et al., retaining their ninefold categorization scheme. Because the HEEDA data base will be used for QSAR prediction, two overall end points are assigned to each chemical for each strain of rat or mouse tested. End points are assigned one of four results: +, −, I, or V. This involves expanding the equivocal rating to distinguish I and V, which enables chemicals with borderline evidence for carcinogenicity (V) to be retained with the positives while eliminating those chemicals with insufficient evidence for determination of carcinogenicity (I). When results are not clear-cut, a comment field from the technical report, which focuses on the experimental problems, is entered. Tumor sites and tissue types are entered as well as route of administration and chemical purity.

(*2*) *The Lijinsky Chronic Toxicity Tests on N-Nitroso Compounds.* In developing his structure–activity data for mutagenicity and carcinogenicity of *N*-nitroso compounds,[8] Lijinsky screened over 100 *N*-nitroso compounds in rats by using uniform test procedures. Many of the compounds tested were chosen in order to elucidate structure–activity relationships in carcinogenicity of these compounds and were synthesized at FCRC and ORNL for the screening tests. Test chemicals were administered orally, usually in drinking water, and all chemicals were tested in chronic studies. Data are entered in the HEEDA data base with carcinogenic potency estimates assigned by Lijinsky (++++, +++, ++, +, 0), sex and strain of test animals, time to death of 50% of animals with tumors, total dose per rat, and target organs.

(*3*) *IARC Monographs.* The IARC panels have reviewed almost 400 individual chemicals or chemical categories. The

Table II. Types of Modeling Programs Available to the Health and Environmental Effects Data Analysis System

modeling method	mathematical methodology	variable type		relative number of independent variables	comments
		dependent	independent		
regression[a]	Best polynomial fit by least squares.	continuous	continuous	small	Regression has been used to predict unknown substance toxicities and various drug activities from measures of lipophilicity, electrophilicity, etc.
National Cancer Institute (Hodes) and Eastman Kodak[b,c]	Test of statistical significance based on difference from population mean.	continuous or discrete	discrete	large	Hodes-Tinker model: Hodes developed this model at NCI for use with large, diverse classes of compounds to predict antineoplastic potential. Tinker incorporated the Hodes algorithm into a system with more complete display capabilities which further reveal the correlations computed. Applications include structure–mutagenicity prediction.
binary partition[d]	Computation of a hyperplane to optimally separate training set compounds into two distinct regions in space.	discrete	continuous or discrete	small	Primarily intended for structurally similar classes of chemicals. Dependent variable can have only two discrete values.
K nearest neighbor[e]	Develops clusters in space based upon minimizing the Euclidean distance among compounds with the same value of dependent variable.	discrete	continuous	small to moderate	Similar in application to binary partitioning except that the dependent variable may have more than two discrete values.
classification analysis method (CAM)[f]	A heuristic method of building a hierarchic classification by repeated partitioning of compound sets according to the cooccurrence of descriptors in compounds with the same value of dependent variable.	continuous or discrete	discrete	large	May be used for very large training sets with many structural features or other independent variables.
BOOLAID[g]	Construction of a hierarchial filter of independent variables using a combination of Boolean logic and analysis of variance.	continuous or discrete	discrete	small	Similar in operation to CAM, but machine time limitations restrict the number of chemical compounds and structural features it can handle.
SIMCA[h]	A pattern recognition technique appropriate for multivariate data sets; in biology, it applies to those problems in which each end point is dependent on several variables. It first seeks to classify carcinogens that are mechanistically similar in their mode of action. Outliers are not necessarily inactive but may be excluded because biological activity is qualitatively different. Within a cluster of biologically similar compounds, a princi-	continuous	continuous	moderate to large	Pilot studies are currently being performed with ecotoxicological data from HEEDA. SIMCA has been identified as a potential addition to the HEEDA library of models.

pal components analysis defines key structural variable and their ranges. Descriptive or dependent variables which have been identified as significant are then related to level of biological activity, such as relative potency.

[a] Martin, Y. "Quantitative Drug Design: A Critical Introduction"; Marcel Dekker: New York, 1978. [b] Hodes, L.; Hazard, G. F.; Geran, R. J.; Richman, S. "A Statistical Heuristic Method for Automated Selection of Drugs for Screening", *J. Med. Chem.* 1977, *20*, 469–75. [c] Tinker, J. F. "Relating Mutagenicity to Chemical Structure", presented on April 23, 1980, as part of the Symposium on Development and Use of Reliable Data Bases for Quantitative Structure–Activity Relationships (QSARs) during the 14th Middle Atlantic Regional Meeting of the American Chemical Society, King of Prussia, PA. [d] Stuper, A. J.; Jurs, P. C. "Classification of Psychotropic Drugs as Sedatives or Tranquilizers Using Pattern Recognition Techniques", *J. Am. Chem. Soc.* 1975, *97*, 182. [e] Kowalski, B. R.; Bender, C. F. "The K-Nearest Neighbor Classification Rule (Pattern Recognition) Applied to Nuclear Magnetic Resonance Spectral Interpretation", *Anal. Chem.* 1972, *44*, 1405. [f] Kulp, Carol S.; Hill, Helen N.; Lefkovitz, David "ETN-4, An Example of Model Building and Test Compound Toxicity Prediction Using BOOLAID and SAMII", University of Pennsylvania, June 16, 1978, EPA Contract No. 68-01-4643, Subcontract No. 104. [g] Lefkovitz, David; Kulp, Carol S.; Hill, Helen N. "ETN-1, Application of BOOLAID Algorithm to the Correlation of Chemical Substructures with End-Point Activity", University of Pennsylvania, Nov 11, 1977, EPA Contract No. 68-01-4643, Subcontract No. 104. [h] Dunn, J.; Wold, S. "Structure Activity Studies by Means of the SIMCA Pattern Recognition Methodology", in "QuaSAR Research Monograph 22"; Barnett, G., Trsic, M., Willette, R., Eds.; National Institute on Drug Abuse: 1978.

chemical categories are usually associated with industrial processes, such as compounds associated with mining beryllium ore. Reviews are published in a series of monographs and have been tabulated in two summary articles and a WHO report.[4,5] These summary publications have made it possible to distinguish three end points for data on cancer in animals reviewed through Volume 16 of the series: sufficient evidence for carcinogenicity (S), limited evidence for carcinogenicity (L), and insufficient evidence for evaluation (I). From volume 17 onward, tighter criteria were applied in the IARC reviews, and chemicals are entered in HEEDA with end points of S or I. In only a few cases were experiments strong enough for negative end points. Althouse et al.[4] provide a summary of results in humans, often utilizing data available after publication of the original monographs. HEEDA entries for human carcinogenicity are taken primarily from this summary publication in order to reflect the addition of later studies.

When reviewers considered chemical categories, they were not always able to delineate well-substantiated results for the individual chemicals comprising the category. When it was possible to differentiate them, an individual entry appears for each reviewed chemical with its CAS number. When data were sufficient only to rate the hazard of chemicals associated with one industrial process, one entry appears for the category. In one example, "beryllium and its compounds" appears as the categorical name, and a CAS number for beryllium is used for the categorical entry.

Mutagenesis. The EPA, under an interagency agreement with ORNL, set up the GENE-TOX program in 1978.[10] Twenty-four panels of experts on various types of genetic toxicology test systems have met to evaluate relevant literature drawn from the Environmental Mutagen Information Center data base for two major purposes: (1) where testing an individual chemical is judged adequate, to assign an end effect value of positive or negative to the chemical for the given test system, and (2) to recommend improvements in the test methods.

Each panel produces a final report at the end of its deliberations that addresses both purposes. End effects values are tabulated in these reports with the experimental data upon which decisions were based. The tabulated data are extracted from the final reports for entry into HEEDA. Certain types of data, called core data, are common to all these reports. Core data include CAS number, chemical name, chemical class assignments, authors of the final report, genetic end point measured, organism identification (species, strain, cell line, sex, etc.), and end effect value.

Other tabulated data types vary according to the test system and are added to the data extracted to form augmented core data (ACD). For example, "germ cell stage tested" is an important piece of data for the heritable translocation, *Drosophila* recessive lethal, and mouse specific locus tests (i.e., tests which involve genetic events in germ cells). "Locus" is required in point mutation tests; "metabolic activator indicator" is required in bacterial and cell culture systems; in cases such as sister chromatid exchange, which is observed by using a number of techniques, "in vivo" vs. "in vitro" is recorded.

Environmental Measurement. Syracuse Research Corporation, under contract to EPA,[22] is reviewing literature related to environmental fate measurements and extracting data that will be stored in HEEDA. These include chemical properties and measurement data, such as

(1) chemodynamic (water solubility, vapor pressure, octanol/water partition coefficient, ultraviolet spectra, dissociation constant, soil adsorption constant),
(2) transport (Henry's law, rate of evaporation from water, bioconcentration factor, soil thin layer and column chromatography),
(3) degradation in various systems,
(4) monitoring data (air, water, and soil when available in the literature reviewed),
(5) field studies.

These data may be used to predict environmental fate and test these predictions as well as to assist in evaluation of exposure for risk assessment. Further, chemical properties may be tapped as a source of chemical descriptors for end effects modeling.

Fish and Wildlife Ecotoxicological Data. Under an interagency agreement between EPA and the U.S. Fish and Wildlife Service (Department of the Interior),[12] the results of ecotoxicological testing performed in various U.S. Fish and Wildlife Service laboratories are being processed for entry into HEEDA. The tests performed in these laboratories were for purposes of screening chemical compounds for animal damage control. Many experimental series of compounds were submitted by industry to the U.S. Fish and Wildlife Service in the course of cooperative research in animal management. Thus, many sets of chemicals from these files have not been registered by the Chemical Abstracts Service. Furthermore, much of this information has been previously unavailable in the scientific literature.

Because the testing of these compounds varied in scope from range-finding data to tests which yield statistically significant results for LD_{50} determination, a house evaluation code has been developed which reflects this variation. This code is assigned by U.S. Fish and Wildlife Service scientists to each test entered into HEEDA.

Files of U.S. Fish and Wildlife Service data presently under development are the following:

(1) Denver Research Center: This has about 4000 compounds representing 5000 tests in about 80 species of birds and mammals. End points measured are lethality and acceptance/repellancy. The tests include dose–response curves for determination of LD_{50}s as well as range-finding studies.[12]

(2) Aquatic Toxicity Files: This comprises about 6000 chemicals in about 15 000 tests of aquatic toxicity to lampreys, carp, and other fish. Many of these tests are range-finding studies.[23-25]

(3) Toxicology Data System: This comprises about 4000 chemicals tested for a variety of effects in birds and mammals.[11]

REFERENCES AND NOTES

(1) Toxic Substances Control Act Chemical Substance Inventory, U.S. Environmental Protection Agency, Office of Toxic Substances, Washington, DC 20460, May 1979.

(2) Bracken, M.; Dorigan, J.; Hushon, J.; Overbey, J. "Chemical Substances Information Network, Volume 1: User Requirements and Systems Development Options", MITRE Technical Report MTR-7558, Contract No. CEQ7A010, June 1977.

(3) Lefkovitz, D.; Hill, H. N.; Kulp, C. S. "System Requirements Analysis for the Chemical Structure and Nomenclature System (CSNS), Final Report", University of Pennsylvania Contract No. EQ8AC027, Sept 28, 1979.

(4) Althousse, R., et al. "An Evaluation of Chemicals and Industrial Processes Associated with Cancer in Humans Based on Human and Animal Data: IARC Monographs Volumes 1 to 20", *Cancer Res.* **1980**, *40*, 1–12.

(5) Tomatis, L., et al. "Evaluation of the Carcinogenicity of Chemicals: a Review of the Monograph Program of the International Agency for Research on Cancer", *Cancer Res.* **1978**, *38*, 877–885.

(6) Griesemer, R. A.; Cueto, C., Jr. "Toward a Classification Scheme for Degrees of Experimental Evidence for the Carcinogenicity of Chemicals for Animals", *IARC Sci. Publ.* **1980**, *27*.

(7) Sontag, J.; Page, N.; Saffiotti, U. "Guidelines for Carcinogen Bioassay in Small Rodents", *DHEW Publ. (NIH) (U.S.)* **1976**, *NIH 76-801*.

(8) Lijinsky, W. "Carcinogenic and Mutagenic *N*-Nitroso Compounds", *Chem. Mutagens* **1976**, *4*, 193–217.

(9) Richmond, S.; Hazard, G. F.; Kalikow, A. K. "The Drug Research and Development Chemical Information System of NCI's Development Therapeutics Program", *ACS Symp. Ser.* **1978**, *No. 84*, Chapter 13.

(10) Waters, M.; Auletta, A. "The GENE-TOX Program", presented on April 23, 1980, as a part of the Symposium on Development and Use of Reliable Data Bases for Quantitative Structure–Activity Relationships (QSARs) during the 14th Middle Atlantic Regional Meeting of the American Chemical Society, King of Prussia, PA. Also, Interagency Agreement No. EPA-IAG-78-D-X0453 between U.S. Environmental Protection Agency and Oak Ridge National Laboratory, Department of Energy.

(11) Menzie, C. M.; DeWitt, J. B.; Walker, C. R.; Bowles, W. A., Jr. "U.S. Fish and Wildlife Service Toxicological Data Retrieval System", in preparation for publication by NTIS.

(12) Walker, C. R. "Evaluation of an Information Retrieval System for Assessment of Toxicological Effects of Chemicals on Fish, Wildlife, and Ecosystem Components", presented on April 23, 1980, as part of the Symposium on Development and Use of Reliable Data Bases for Quantitative Structure–Activity Relationships (QSARs) during the 14th Middle Atlantic Regional Meeting of the American Chemical Society, King of Prussia, PA. Also, Interagency Agreement No. 14-16-0009-79-986 between U.S. Fish and Wildlife Service and the U.S. Environmental Protection Agency.

(13) See, for example: "Biological Correlations–The Hansch Approach", *Adv. Chem. Ser.* **1972**, *No. 114*.

(14) Wassom, J. S. "The Storage and Retrieval of Chemical Mutagenesis Information", in "Progress in Environmental Mutagenesis"; Alacezic, M., Ed.; Elsevier/North Holland: Amsterdam, 1980; pp 313–330.

(15) Milne, M.; Lefkovitz, D.; Hill, H.; Powers, R. "Search of CA Registry (1.25 Million Compounds) with the Topological Screens System", *J. Chem. Doc.* **1972**, *12*, 183.

(16) BASIC Substructure Search/Fragment Search Dictionary, Part I: Fragment Type Sequence. Basle, Feb 1977. Part II: Fragment Number Sequence, June 1977.

(17) Van Meter, C. T.; Goldschmidt, E. N.; Milne, M. CIDS No. 6, Handbook of CIDS Chemical Search Components, Status Report, University of Pennsylvania, Philadelphia, Dec 1968.

(18) CO70 Inquiry User Manual, distributed by Chemical Abstracts Service Central Documentation, July 8, 1977.

(19) Smith, E. G. "The Wiswesser Line-Formula Chemical Notation"; McGraw-Hill: New York, 1968.

(20) Miles, P. C.; Sunden, A.; Boaler, H. J. "Instructions for the Selection and Presentation of Data for the International Register of Potentially Toxic Chemicals", Kratel Documentation and Research Centre Contract No. G/CON/79/05-UNEP/IRPTC, 1979.

(21) Toxicology Data Management Systems, prepared by Division of Toxicology Data Management Systems, Department of Health, Education, and Welfare, Food and Drug Administration, National Center for Toxicological Research, Jefferson, AR 72079, Feb 12, 1980.

(22) Cooperative Agreement for Development of an Environmental Data Base, Grant ID No. CR806902-01-0 between Syracuse Research Corporation and Office of Pesticides and Toxic Substances, U.S. Environmental Protection Agency.

(23) Applegate, V. C.; Howell, J. H.; Hall, A. E., Jr.; Smith, M. A. "Toxicity of 4,346 Chemicals to Larval Lampreys and Fishes", *U.S. Fish Wildl. Serv., Spec. Sci. Rep.–Fish.* **1957**, *207*, 1–157.

(24) Loeb, H. A.; Kelly, W. H. "Acute Oral Toxicity of 1,496 Chemicals Force-Fed to Carp", *U.S. Fish Wildl. Serv., Spec. Sci. Rep.–Fish.* **1963**, *471*, 1–124.

(25) MacPhee, C.; Ruelle, R. "Lethal Effects of 1,888 Chemicals upon Four Species of Fish from Western North America", University of Idaho Forest, Wildlife, and Range Experiment Station Bulletin No. 3, 1969, pp 1–112.