more than one industrial application. The second requirement will ensure more meaningful test results. If a company's entire product line had only one end-use, the coding of its products would be oversimplified, and potential problems may escape notice.

Because we cannot determine in advance the specific chemicals or categories of chemicals which will be reported via the classification system, our major criterion for selecting test chemicals was that they be commercially marketed and not restricted to research applications. In addition, we have selected chemicals used in representative industrial applications. To the extent possible, we have selected chemicals handled by two or more companies. This overlap among chemicals will enhance the test of the classification system by indicating the correlation between companies in interpreting the use of the classification system. Any changes in the classification system deemed necessary as a result of the test will be incorporated into the system.

CONCLUSION

Development of this classification system addresses a vitally important component of EPA's functions under the TSCA: the collection of chemical use data in a form that permits subsequent evaluation and analysis. The end goal of classification system development is a data collection instrument which will allow manufacturers and processors of industrial chemicals to report codified chemical use data to EPA.

LITERATURE CITED

(1) U.S. Congress Senate, Toxic Substances Control Act, Pub. L. 94-469, 94th Congress, 2nd Session, S. 3149, Oct 11, 1976.
(2) U.S. Office of Management and Budget, "Standard Industrial Classification Manual", U.S. Government Printing Office, Washington, D.C., 1972.

CHEMFILE: An In-House Information System for the Chemical Indexing of *Abstracts on Health Effects of Environmental Pollutants* (HEEP)†

WILLIAM GRAHAM*
BioSciences Information Service, Philadelphia, Pennsylvania 19103

Received June 1, 1977

The inclusion of a Chemical Abstracts Service (CAS) Registry Number index in HEEP has led to the need for a special database designed to link substance names with their appropriate CAS chemical compound Registry Numbers. Begun in 1968 as a tape record with batch mode updates and few record modification capabilities, the information system, which we call CHEMFILE, has evolved to its current form of disk storage with on-line access for file maintenance.

INTRODUCTION

The BIOSIS CHEMFILE is an information system developed to assist in the assignment of Chemical Abstracts Service (CAS) Registry Numbers to substance names occurring in the biological literature. Currently only 23 000, or approximately 10% of articles reviewed by BIOSIS, have substances uniquely identified but the commitment to this technique and the system needed to perform it provide BIOSIS with an operating model that allows us to learn the requirements that must be met to allow this indexing for other BIOSIS products. We now include CAS Registry Numbers as index terms in the printed version of *Abstracts on Health Effects of Environmental Pollutants* (HEEP) and the tape record of abstracts and citations dealing with drug toxicity now sent to the National Library of Medicine for inclusion in the TOXLINE database.

START OF CHEMFILE

In 1972, BIOSIS began publication of HEEP, a subset of abstracts and citations covered by BIOSIS and the National Library of Medicine. The project specifications required that a separate index of Registry Numbers for substances mentioned be included. CAS has cooperated in this work by providing Registry Numbers which could not be located in the printed indexes to *Chemical Abstracts* or other Registry Number sources.

Since the number of different substances encountered in HEEP was relatively small and the frequency of occurrence high, a file which BIOSIS now calls CHEMFILE was developed to save having to locate any given Registry Number more than once. A file similar in nature begun for an earlier project was built upon as new substances were identified. A record was prepared for each substance which contained the following types of data:

BIOSIS accession number
CAS Registry Number
Molecular formula
CAS Type 1 Name
Synonyms such as trade, generic, and systematic names
Wiswesser Line Notation

The serially assigned BIOSIS accession number tied all of the information together. It was used, rather than the CAS Registry Number, since substances and publication items could be processed further while the Registry Number was being located. Further, since not all chemical entities impacting on toxicology studies are registerable, we wanted to retain what information we had—to avoid looking for a Registry Number more than once if it did not exist.

CHEMFILE STRUCTURE

Each record, originally on magnetic tape, contained the accession number, a code for data type (10 for Registry Number, 20 for molecular formula), and the content of the record. The tape record was fully listed periodically to provide a hard copy for reference. The principal tool derived from the file was an alphabetical listing of all substances encountered. A third output was a numerically ordered listing of all Registry Numbers in the file. This was checked manually to assure that no Registry Number would be entered more than once.

The file was updated in large batches by inputting data from keypunch cards. Various validation checks, such as check digit

† Presented at the 11th Middle Atlantic Regional Meeting of the American Chemical Society, Newark, Del., April 20–22, 1977

* Present address: AMP Inc., Harrisburg, Pa.

calculations on accession number and Registry Number and tests for proper card format, were performed at input. The following were problems encountered with this system:

1. Errors were detected only when a batch run was made. The run was halted and the cards and error messages were returned, the cards were corrected, and another run was scheduled.
2. Since the listings were lengthy, they were run less frequently than updates. As a consequence, it was difficult to know what was on the machine record between listings.
3. The process for correcting names was cumbersome since the incorrect record had to be keyed on a card exactly as it appeared in the file (to locate the incorrect record) and again with the correction made. This required rekeying correct information and allowed for possible introduction of new errors.

REDESIGN

Recognizing these difficulties, a new system for the CHEMFILE was designed and became effective in March 1976. It offers better access to the machine-readable record, improved correction capabilities, and error detection while keyboarding.

The computer in use at BIOSIS is an IBM 370/145. Two IBM software packages, CICS and VSAM, are used to control terminal communications and disk access. IBM 3277 terminals are used for both input and searching. The file now contains information not previously stored (date of record creation, date of last change, and employee number of the operator) as well as all previously recorded data.

The principal record identifier is still the BIOSIS accession number and is contained in all records dealing with a particular substance. The following are the data elements for substance information on the file:

7010 CAS Registry Number
7020 Molecular formula
7030 CAS Type 1 Name
7040 Synonyms
7050 Wiswesser Line Notation

Each data type may have up to 99 modifications (mods). For example, the synonym data element may contain 99 different synonyms number 7040-01, 7040-02, . . ., 7040-99. Further flexibility has been incorporated to permit 99 submodifications (submods) of each synonym or Type 1 Name. Presently 7040-01-01 is the standard CHEMFILE designation for the first synonym for a given substance; 7040-01-02, the sortkey for the alphabetized listing; 7040-01-03, the BIOSIS Previews version of the name if these forms are later added to the file; etc.

RECORD CREATION AND UPDATES

Use of the CHEMFILE maintenance system requires that the operator know the accession number, data type, and mod and submod numbers which will serve as record identifiers for the new information or to retrieve a record already in the system. Upon signing on, the operator is presented with a video display, called the *selection* screen, which is formatted to accept all numerical data needed to identify the record. When this information is entered, the system returns a *record* screen which has three fields. The upper field displays the data keyed on the selection screen, a written description of the data type selected, and all other information automatically stored for each record. This additional information—date of creation and operator, for example—will have no values recorded if the record is being input for the first time.

The second field is the data field and may contain up to 1000 characters. This field is used to display or key one data element, such as the Registry Number or a substance name. The third field displays a menu of the functions performed by the five program function keys on the terminal. These deal mainly with paging within a record and between records.

The entire screen on the terminal, therefore, displays information pertaining to only one record in the file. Paging forward from one synonym of a substance to the next will display a new screen of information for the next synonym. To add a new synonym, the operator must call for the next unused mod number or use a program function key which will display a screen complete with the next highest unused mod number.

New Registry Numbers and substance names are keyed in, one to a screen and entered individually. Each time a record entry is made the CHEMFILE system displays the original screen with record creation and operator data and the information input.

Invalid data type, mod, and submod numbers are automatically rejected as incorrect. The system also checks BIOSIS accession numbers and CAS Registry Numbers at input by performing a check digit calculation. Any corrections that need be made, whether to such a number or to a name, are made by positioning the cursor under the incorrect character, changing it, and reentering the corrected name. A record which is incorrect due to having an extra character or characters in the middle of a correct line may be corrected by a delete function which removes the extra character and closes up the line behind it. Extra characters may be inserted without disturbing any information before or behind them. The maintenance system, itself, removes extra blanks when more than one are keyed between text words.

CHEMFILE REPORTS

The following printed reports are available from the system.

1. Data base maintenance report: This is produced nightly and lists any activity on the file during the work day. This report is used for proofreading input records.
2. Registry Number listing: a numerically sorted listing to check whether newly located numbers are already in the system.
3. Alphabetical listing: a printout of CAS Type 1 Names and all synonyms in the file which is sorted by use of a generated sort key.
4. Accession Number-ordered listing: This large listing may be run to list all records on the file or to list all records within a given accession number range. This is used primarily to look up records away from the terminal.

SEARCHABLE FILE

The system described, thus far, is comprised of the programs, control systems, and data file used to build and maintain the CHEMFILE at BIOSIS. Access to it is limited to those authorized to effect changes to any data on record. It is not available to searchers or indexers for use in their tasks and would be inappropriate if it were so, since it has no search capabilities and cannot show all information about a given substance in a unified display. It has been designed for the input of selected data elements about substances and later verification that those records have been added.

The indexing and searching staff do, however, need access to current versions of the CHEMFILE in order to assign numbers to new substances or to build strategies based upon known synonyms and systematic names. To this end, the CHEMFILE is converted to files searchable under the IBM software package, STAIRS, for on-line retrieval. The development of current programs for displaying and searching records of chemical substances is based upon experience gained with an early version of the CHEMFILE loaded onto STAIRS three years ago.

While some of the development of programs and rules to describe exactly how systematic names will be searched is only

now nearing completion, the file is loaded under STAIRS and available to staff members who may sign directly on to the file or who may access it during a search session with other BIOSIS on-line files. Since no portion of the file may be altered during these sessions, the file may be used by many staff members rather than a restricted few and the users need not learn any command language other than STAIRS to use it.

A given substance is retrieved through constructing the known name and displaying the resultant hits. The logical operators AND, OR, NOT, and ADJ may be used to construct searches based upon name fragments. Search results may be sorted upon Registry Number if desired, and results may be more selectively chosen by specifying Accession Number or Registry Number ranges. The record displayed lists the information in file about a substance. Verbal descriptions of the data type head each portion of the display and each synonym is listed on a separate line. STAIRS also offers the user immediate copies of a retrieved document from a remote printer, or overnight printing of entire search results.

Database Development in a Regulatory Agency†

MARILYN C. BRACKEN*
The MITRE Corporation, McLean, Virginia 22101

IRVIN J. WEISS
U.S. Consumer Product Safety Commission, Bethesda, Maryland 20207

Received April 5, 1977

A general discussion of the history and problems associated with the collection of information and the development of a database in a regulatory agency is presented. The proceedings of the U.S. Consumer Product Safety Commission for obtaining chemical formulation information for specified consumer products is discussed. Guidelines for database administrators faced with data collection activities in a regulatory agency are provided.

INTRODUCTION

The Consumer Product Safety Commission (CPSC) is responsible for programs that reduce the hazard of human injury from chemical consumer products. To make scientific conclusions and value judgments about the safety of a product, complete information on ingredients in the product is required. This paper describes the nature of proceedings of the U.S. Consumer Product Safety Commission to obtain chemical formulation information for specified consumer products, including (1) the issuance of a Special Order to obtain such information and (2) the defense against legal actions brought by a trade association to obtain a preliminary injunction to prevent the Commission from collecting this information. Recommendations for others involved in data collection and database development are presented.

BACKGROUND

Unlike the Food and Drug Administration and the Environmental Protection Agency, the Commission has no preclearance requirements in its laws which obligate manufacturers to seek approval for marketing their products. However, the Commission does have the authority (which it has not yet implemented) to prescribe procedures for the purpose of ensuring that the manufacturer of any new consumer product furnish notice and a description of such product to the Commission before its distribution in commerce. The Commission may also, by rule, require any manufacturer of a consumer product to provide to the Commission performance and technical data that relate to performance and safety as the Commission determines necessary.

In 1973, the Commission decided that in order to respond to its Congressional mandate to protect the public from unreasonable risk of injury, and for effective enforcement, product ingredient information for chemical consumer products was required. A project was initiated to collect formulation information for selected consumer products.

USES TO BE MADE OF THE DATA

The purpose of collecting formulation information on chemicals in consumer products is to provide the Commission with means of ascertaining the potential or real hazards to consumers due to exposure. The Commission receives petitions, consumer complaints, notices of product defect, and Congressional inquires. It is asked, for example, to consider the banning of a class of product(s) containing a certain chemical. With no detailed knowledge about the ingredient data of such products, it is extremely difficult to predict or evaluate the hazards involved.

There is a need to be able to characterize types of consumer products. For example, what ingredients would one find in tile adhesives or in dishwasher detergents or in any particular category of chemicals. With knowledge of formulation information the Commission can: (1) identify chemicals that are toxic, carcinogenic, mutagenic and teratogenic, (2) look at particular classes of compounds and develop a standard, or (3) address generic standards.

The Commission through its interagency committee activities receives notification of chemicals suspected of being carcinogenic and/or documented to be cancer-causing agents in animals and man. Knowledge of the existence of such chemicals in products that are under the Commission's jurisdiction is critical, particularly if a ban is under consideration. Plans for recall must be developed, as well as economic analyses regarding impact on the affected industry of generating a substitute for the chemical.

In addition, the formulation information will assist the Commission in implementing Section 13 of the Consumer Product Safety Act. This section is designed to provide the

† Presented at the 39th Annual Meeting of the American Society of Information Science, San Francisco, Calif., Oct 4–9, 1976.