59034244

ATENB/ATMDS ENVIRON

7 (7). 1974 749-754

HUCKABEE, JW/

MOSSES SENSITIVE INDICATORS OF AIRBORNE MERCURY POLLUTION/
GRASSES VEGETATION

01008 04500 06504 07504- 07506 10010 10059 10069 13010-
22506- 37015* 51519- 51526-

1100 21600 25305

Figure 5.

TOXICITY OF IODINE USED AS A DISINFECTANT

IODIDE/

IODINE/ AND CC37008 DISINFECT/VECTOR CONTROL AND CC22504 PHARM TOX
IODINATED CC39500 DISINFECT/STERILIZATION CC22506 ENVIRON TOX
IODO/ CC22508 VET TOX

Figure 6.

case, a combination of natural language keywords for Iodine and subject categories for Toxicity and Disinfectants was used. It was necessary in this search to have a parameter for each concept in order to minimize the identification of irrelevant material. The CROSS or concept codes have again been used for Disinfection and Toxicity.

Figure 7 shows a sample of an item retrieved from this search. Many descriptors have been added to enrich the

75001344

BVJOA/BR VET J

156 (2-3). 1974 145-156

WILLINGER H/ THIEMANN G/

CRITICAL ASSESSMENT OF STERILANT AGENTS IN THE FIELD OF VETERINARY HYGIENE/

HUMAN ANIMALS FORMALDEHYDE CHLORINE COMPOUNDS PHENOL DERIVATIVES IODOPHOR TOXICITY

10058 10060 10069 22501 22504- 22508- 26502 38002- 39500*

85150 86215

Figure 7.

authors' title. This item resulted from a match between the keyword Iodophor and two CROSS codes, Disinfection and Sterilization represented by the number 39500 and Veterinary Toxicology represented by the number 22508. The Bio-Systematic codes, shown on the last line of the item, refer to general vertebrates and human studies.

The manner in which a large broadly based machine-readable file can be used to help to meet the needs of scientists concerned with environmental problems has been described. The potential for exploitation of such a file is almost limitless and can contribute significantly toward helping the scientific community to deal with some of the problems arising in a relatively specialized subfield of biology.

Bridging and Interlinking the Information Resources†

DALE C. MYERS, JOYCE A. RATHBUN,* FRED A. TATE, and DAVID W. WEISGERBER

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received November 5, 1975

The American Chemical Society through its Chemical Abstracts Service Division plans to establish a Regulatory Registry Service (RRS), based on the CAS Chemical Registry System, to uniquely identify substances in publicly available government agency files and thereby to provide a potential interfile link for substance-oriented data in these files. Use of the RRS will avoid duplication of effort in producing and compiling information and will improve access to substance-related information contained in existing publicly available government files.

INTRODUCTION

Today's national and international concerns—energy, natural resources, ecology, food, health, consumer protection, urban development, transportation, etc.—are heavily overlapped and tightly interdependent. Likewise, governmental efforts directed at solving these problems have led to a proliferation of interrelated and often overlapping legislation and regulation at all levels of government, both in this country and internationally. Such legislation frequently assigns responsibility for compliance to more than one agency. For example, of a sample set of 40 U.S. federal regulatory laws, 13 were found to deal with the EPA, 6 with the FDA, 5 with the Consumer Product Safety Commission, 4 with the Drug Enforcement Administration, 10 with the Materials Transportation Bureau, and 16 with other agencies. These add up to more than 40 because of the substantial overlap in agency responsibility in just this one small sample.

Currently, many different government agencies or even different offices within the same agency share responsibility or have similar responsibilities for related regulation. Of course, these interrelationships are not limited to the federal agencies. The states and many municipal and other local government activities also must deal with the same problems. The Ohio Environmental Protection Act, just one of many possible examples, includes nine regulations on water, seven on air, and one on solid waste.

The probable overlap of agency interests has vast dimensions when one considers that the multiplicities of regulatory overlap which exist within the U.S. are for just one country. Other countries have similar internal concerns and overlaps. And the flow of commerce and the universality of social problems throughout the world add still other layers of overlapping interest and regulation.

To deal effectively with this overlap of government interest as it relates to the accessing of substance-related information accumulated by government agencies, the American Chemical Society (ACS) through its Chemical Abstracts Service (CAS) Division plans to develop and implement a Regulatory Registry Service (RRS) which will be described in the balance of this paper. To simplify the following discussion, these comments will concentrate on dealing with the files of U.S. federal agencies. Also, although a fully implemented RRS would deal with the files of all U.S. federal agencies, initial attention in

† Presented in symposium on "Information Requirements Resulting from Environmental Impact Laws", Division of Chemical Information, 170th National Meeting of the American Chemical Society, Chicago, Ill., Aug 27, 1975.

* To whom correspondence should be addressed.

implementation would focus on selected files.

SUBSTANCE INFORMATION IN REGULATORY AGENCY FILES

Much of the regulatory legislation related to current social problems concerns natural resources, manufactured materials, and wastes, all identified as substances. The range of public concerns related to substances is extensive: sources of raw materials, manufacture, packaging, labeling, potency, storage, purity, misuse, accidental release, transportation, etc. And each of these concerns interfaces within and among local and national governments and private enterprise worldwide.

As the various agencies, offices, and departments fulfill their assigned duties (support research, gather data, enforce regulations, provide public reports, etc.), they amass a great deal of information about chemical substances. Currently there is no generally accepted means for identifying such information associated with any given substance contained in just one agency's files to say nothing of the possibility of linking related information in the files of two or more agencies.

One agency's files are often partially or completely duplicated for use at multiple sites, and the files of different agencies frequently include the same or related data, with each agency's file tailored to that agency's specific needs. It is not the overlap which is costly, but rather *unrecognized* overlap which leads to unnecessary duplication of data gathering, file building, and system maintenance. Perhaps even more important in dealing with pressing social problems is the fact that unrecognized overlaps and associations among existing data resources do not permit the use of combinations of these resources in seeking the most desirable solutions. In fact, the inability to identify the existence of needed data or to promptly establish interlinks to those data can prevent rational handling of a problem which could otherwise be solved.

As has been stated in a report by the U.S. Comptroller General,[1] government agencies are seldom required by law to know what is in the files of other agencies. This is not surprising, for until the recent development of the CAS Chemical Registry System, no reliable means of providing consistent, nonvariant referencing of substances was available.[2] Now, however, with the CAS Registry being full operational and widely accepted by worldwide scientific and technical communities, this proven system can provide the basis for supplying the necessary nonvariant identifiers reliably and economically.

THE REGULATORY REGISTRY SERVICE

Chemical Abstracts Service is now planning to initiate a Regulatory Registry Service designed to supply reliable, nonvariant substance identifiers which can be used for referencing substances both in computer-readable files and in printed works. The content of the RRS files will initially come from information made available to the public by U.S. federal agencies. The system will be made available for government and nongovernment use. Such use will not be a part of the public record, and no proprietary data will be incorporated into RRS files. It is anticipated that RRS services will be useful to legislative bodies; federal agencies; national, state, and local governmental operations here and abroad; international organizations; business; industry; commerce; etc. The foundation of the RRS will be the CAS Chemical Registry System.

CAS Chemical Registry System. The Chemical Registry System[3] is based on an algorithm that generates a unique and unambiguous computer-language description of a chemical substance's two-dimensional structure and stereochemical details. The algorithm then assigns a computer-readable and computer-checkable code called a Registry Number to each substance. Each number designates only one substance, and this system of identification is independent of nomenclature.

The names of chemical substances, even when systematically derived, have not proven practical for unique identification of substances in information-handling systems. For many substances it is possible to derive two or more completely systematic but different names. Often these names are so dissimiliar that even a nomenclature expert would have difficulty in recognizing that each name identifies the same substance. In addition, many chemical substance names are so long and complex that using them in daily communication or handling them manually in files is inconvenient. One substance in the CAS files, for example, has a systematic name containing 3483 characters. Thus, less systematic or nonsystematic names are often substituted. A nonsystematic name, however, is not a reliable substance identifier since such names often vary from file to file and sometimes even within a single file.

There is no intention here to propose Registry Numbers as a *replacement* for names. Names are, and always will be, useful substance identifiers in individual files and for many forms of communication about substances. Registry Numbers, which are meaningless in themselves, are designed only to facilitate information handling by providing a single, universal means of routinely identifying substances.

Regulatory Registry. Interlinking substance-related information in a wide variety of dispersed information collections such as government agency files requires, first of all, a uniform means for identifying each substance. The planned RRS will employ CAS Registry Numbers and similar identifying numbers to provide unique identification for each substance.

The RRS will contain substances regulated or identified with public concerns, each identified by a CAS Registry Number or similar Registry-like code. Substances in the RRS will be drawn from publicly available files and references such as the Federal Register, Requests for Proposals (RFP's), the Congressional Record, and other publicly available publications. Linked to each substance will be a code referencing the files that contain information about that substance, as well as pertinent legislative references. By comparing the Registry Numbers of substances listed in each agency's files, the RRS would make it possible to identify overlap among the files of the participating regulatory agencies. Provided with this overlap information, the regulatory agencies could be able to avoid much wasteful duplication in generating and using information. Consequently, responsiveness to public needs should be improved through more timely application of existing information and the public should gain a greater return on its cumulative investment in the available information, both public and private.

The CAS Registry Number is a reliable substance identifier and used in conjunction with source identifiers in the RRS would make it possible for government agencies and, if desired, private companies to be aware of substance information available in various agency files. The potential value to the private sector is important because while government files contain much substance-related information, new substances and new research about substances come mainly from the private sector. It would often be useful for an individual organization to tap into publicly available information related to its interests. The proposed RRS would provide the possibility for such interlinking between the files of government and the private sector. The RRS could also be used to simplify the flow of government-required information from industry and help in identifying related information in the published scientific and technical literature.

Characteristics of RRS. RRS would permit simplified interlinkage of substance-oriented information found in widely separated files. The information itself would not be gathered

Table I. Size of Example Files

File	No. of entries
EPA Pesticide File	1455
HEEP Chemical Index Guide	3958
PHS-149 (compounds tested for carcinogenic activity)	3632
NIOSH (suspected carcinogenic compounds)	804
Total entries	9849

Table II. Overlap of Example Files

Substances found in	No. of substances
One file only	5481
Two files only	1342
Three files only	492
Four files	52
Total No. of different substances	7367

together. It would remain in the various agency files but would be identified so that needed information could be reliably located. By linking, but not physically combining, separate files, the advantage to each agency of having its own files would be retained. The use of CAS Registry Numbers only to provide a means to link different files would also mean that each agency could retain existing means of accession to their own files (e.g., by agency-assigned name or by agency-assigned accession code).

Many files, both public and private, contain confidential information. While the RRS would link files, the existing confidentiality in files would be retained since the RRS would contain only substances and not associated data about these substances.

While the RRS will deal only with substances and not contain related information, it could be used to access a substantial body of regulatory literature which is not directly substance related. It could, for example, lead to information about a certain medical device through a reference to the substances used by the device, measured by the device, included in the device, etc.

The CAS Registry System has proven itself capable of handling the large number of substances that would likely be included in a Regulatory Registry System. The CAS Registry System has been in operation since 1965, and in this time approximately 3.3 million substances have been registered. For these 3.3 million substances, there are about 5 million names and 18 million access points. Currently, Registry Numbers are used to reference substances in the CAS data base. In addition these identifying numbers are employed in a wide variety of handbooks, reference works, and journals, e.g., *Journal of Organic Chemistry*, *Inorganic Chemistry*, "USAN 10 (United States Adopted Names) and USP (United States Pharmacopeia) Dictionary of Drug Names," *Abstracts on Health Effects of Environmental Pollutants* (HEEP), and the National Library of Medicine's TOXLINE/CHEMLINE system (an on-line computer-based toxicology information service).

To show the effectiveness of using Registry Numbers to determine overlap between files, CAS compared several files in which substances already have CAS Registry Numbers. As shown in Table I, these files contain a total of 9849 entries. A comparison of unique Registry Numbers indicated considerable overlap among the files (see Table II).

OPERATING THE RRS

CAS plans to include in RRS not only fully characterized substances but also those under regulation in the generic sense. RRS will be operated as a profile in conjunction with the CAS

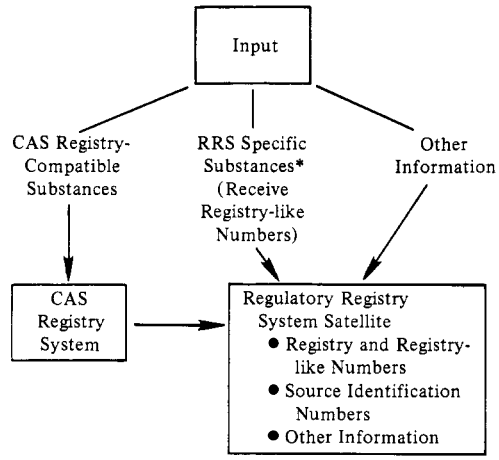

Figure 1. Operation of the Regulatory Registry Service.

Registry System (see Figure 1). Those substances which are defined in a way consistent with CAS Registry operations will be processed through the CAS Registry System where the existing Registry Number will be retrieved or a new Registry Number assigned. Some substances in RRS, those which would normally not be included in the CAS Registry, will be included in RRS and will receive Registry-like numbers, identification numbers which look exactly like Registry Numbers in format and appearance.

In implementing the RRS, CAS proposes to begin with some selected files from government agencies. Once the basic system is in operation and providing services, we will begin system extension. Several types of service will be possible with the RRS. First, RRS would provide registration of substances and maintenance of the files to keep them current. RRS would provide overlap detection on a set schedule. RRS could also identify, for people who have information about a given substance, the sources of other files containing the same substances—not for the purpose of supplying them with information, but for identifying that the possibility for information exists. CA Index Names and synonymous names can also be provided if they are available in the CAS Registry System. One-time search of the files could also be provided. For example, if a proposed piece of legislation dealt with a certain substance, the Congressional Research Service of the Library of Congress would be concerned about which agencies were already regulating or collecting information about that substance. At the present time, finding this information entails a substantial amount of effort. With the help of RRS much of this effort could be avoided.

We anticipate that the eventual audience served by RRS would be legislative agencies at all levels (local, state, and federal), U.S. legislative services, international organizations, industry, and individual workers in science and technology. The initial audience will likely be the U.S. Federal Government but others will be added as the service grows.

REFERENCES AND NOTES

(1) Comptroller General Report: "Federal Environmental Data System (B-177222)," Report to the Subcommittee on Fisheries and Wildlife Conservation and the Environment, Committee on Merchant Marine and Fisheries, House of Representatives, by the Comptroller General of the United States, General Accounting Office.

(2) EPA Order 2800.2 of 27 May 1975. This order establishes the policy, definition, procedures, and responsibilities related to the use of Registry Data from the Chemical Abstracts Service Division (CAS) of the American Chemical Society in ADP systems containing data/information on specific, definable chemical substances. Any computer-based agency scientific and/or administrative system currently in use or being planned

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.