

CONCLUSIONS

In most cases where only single keywords are being used for retrieval of articles in a broad area of interest, maximum success will be obtained by a manual search. It should also be recognized that where a number of nonspecific terms must be searched which do not appear as entries per se in the manual indexes (e.g., for certain chemical reactions or mechanisms involving many different compounds), it is possible that computer searching would yield more references and almost certainly be more cost effective. Our approach to searching the chemical literature involves the use of on-line systems as an efficient starting point upon which to build an information base. Major reliance is still placed on manual searching and includes particularly the analysis of individual bibliographies as provided by selected articles. However, with the recent on-line availability of supplemental data bases such as *Science Citation Index*, dependence on manual methods for complete bibliographic retrieval can be further reduced. The advent of new tape services by Chemical Abstracts Service (e.g., CA Subject Index Alert) which allow for free-text searching of

abstracts should also close the gap considerably between the overall success of computer-based vs. manual searching.

ACKNOWLEDGMENT

Support in manual searching and analysis of data by Ms. D. Christopher is gratefully acknowledged. The work performed in this study was partially supported by Subcontract No. 4481 with the Oak Ridge National Laboratory, Oak Ridge, Tenn.

LITERATURE CITED

- (1) R. E. Buntrock, "Searching *Chemical Abstracts* vs. *CA Condensates*", *J. Chem. Inf. Comput. Sci.*, **15**, 174-176 (1975).
- (2) B. C. Prewitt, "Searching the *Chemical Abstracts Condensates* Data Base via Two On-Line Systems", *J. Chem. Inf. Comput. Sci.*, **15**, 177-183 (1975).
- (3) J. S. Buckley, "Planning For Effective Use of On-Line Systems", *J. Chem. Inf. Comput. Sci.*, **15**, 161-164 (1975).
- (4) C. J. Michaels, "Searching *CA Condensates* On-Line vs. the CA Keyword Indexes", *J. Chem. Inf. Comput. Sci.*, **15**, 172-173 (1975).
- (5) "Information Tools 1976. Book One," Chemical Abstracts Service, Columbus, Ohio, 24 pp.

Building a Chemical Ingredient Data Base for Industrial and Consumer Products[†]

WENDY L. BYER,* HERBERT B. LANDAU, M. LYNNE NEUFELD, and HARRY ROSENTHAL

Auerbach Associates, Inc., Philadelphia, Pennsylvania 19107

Received March 23, 1976

Data bases containing the chemical ingredients of over 100 000 trade name industrial and consumer products have been compiled for the National Institute for Occupational Safety and Health and the U.S. Consumer Product Safety Commission. The methods for obtaining and compiling the ingredient information have relied on computer assistance for standardizing data to preferred formats, for generating individual product ingredient requests and monitoring the status of the response, and for controlling the quality of the information entered into the data base.

INTRODUCTION

In order to fulfill their assignments as guardians of occupational and public health, two government agencies have elected to conduct large-scale surveys to identify chemical compounds to which workers in industry and consumers in the home are routinely exposed. The end result of both projects is a machine-readable data base containing the chemical ingredients of trade name products. This paper reviews the methodology of ingredient data collection and processing procedures, which are similar for both projects, and comments on the differences in scope and data control between the projects.

The National Institute for Occupational Safety and Health (NIOSH), of the Department of Health, Education and Welfare, is required to determine tolerable levels of exposure to hazardous chemicals in industrial environments and to draft recommendations concerning proper use of these chemicals. Before doing so, NIOSH is attempting to discover the incidence of chemical exposures across American industry, such incidences to be reported in terms of type of industry, occupational group, and size of industrial facility. Therefore, NIOSH conducted the National Occupational Hazard Survey (NOHS), comprising site visits to approximately 5000 rep-

resentative industrial facilities and recorded, as occupational exposures, both specific chemical compounds and finished trade name products being used. Because 75% of the data has been reported as exposures to trade name products, it is necessary to reduce these trade names to their chemical components so that NIOSH can enumerate worker exposures to specific chemicals.

Under the Consumer Product Safety Act, the U.S. Consumer Product Safety Commission (CPSC), an independent agency with both investigatory and regulative authority, is responsible for reducing the risk of human injury from chemical consumer products. Trade name products investigated in the CPSC survey were selected at random from among 33 consumer product categories of the National Electronic Injury Surveillance System (NEISS). These categories were chosen and prioritized by CPSC on the basis of injury data reported through the NEISS system. The data compiled on the reported chemical compounds will serve as the basis for monograph development. Each monograph will treat a specific chemical compound and summarize published data on the chemical, biochemical, and biological properties of the compound.

OVERVIEW AND SCOPE OF THE PROJECTS

Both the 33-month NIOSH project, initiated July 1973, and the 24-month CPSC project, initiated July 1974, have a common objective: to accurately and specifically define the chemical ingredients of individual trade name industrial and

[†] This work was performed under National Institute for Occupational Safety and Health Contract No. HSM-99-73-67 and Consumer Product Safety Commission Contract No. CPSC-C-74-218. This paper was presented at the 10th Middle Atlantic Regional Meeting of the American Chemical Society, Philadelphia, Pa., Feb 24-25, 1976.

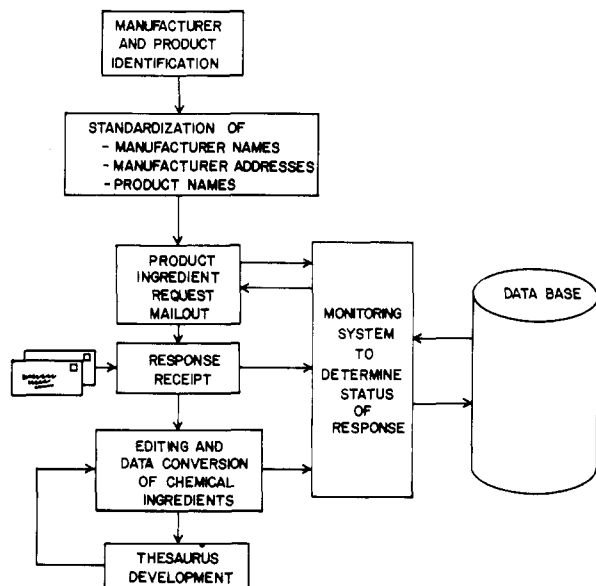


Figure 1. Work flow of procedures.

consumer products. The methodology for accomplishing this involves five operations: (1) product and manufacturer identification and verification, (2) ingredient data collection, (3) response status monitoring, (4) chemical edit and vocabulary control of product ingredient information, and (5) data base compilation. Figure 1 illustrates the relationship of these five operations.

Survey procedures for both projects incorporate several notable technical features, such as computerized editing routines for quality control, computer generated manufacturer address and trade name request labels, and a computerized bookkeeping system for monitoring the status of ingredient responses. The final product is a machine-readable data base that relates chemical ingredients to trade name products and their manufacturers. The data base can be used as input to a generalized report generator, rather than to a specific printed directory.

Several differences exist between the two projects. Input data for the NIOSH project, which concerns industrial products, were collected through a field survey, while input data for the CPSC project, aimed at consumer products, were collected from published sources. Differences also exist in the volume and precision of data collected. The NIOSH project covers approximately 86 000 products manufactured by over 10 000 different companies and requires precision of ingredient reporting at the 1% level. The CPSC project covers approximately 20 000 products manufactured by 1700 companies and requires an ingredient reporting precision at the 0.1% level. In addition, vocabulary control is applied at different stages to the ingredient data received from manufacturers.

PRODUCT AND MANUFACTURER IDENTIFICATION AND VERIFICATION

Product and manufacturer identification procedures differed between the NIOSH and CPSC projects owing to the use of different data sources on each project. The source of NIOSH product and manufacturer names was the NOHS survey data, while CPSC products and manufacturers were identified from the published literature. The NOHS data employed in the NIOSH effort required editing to eliminate synonyms prior to use, as the survey technique employed by the Government allowed for multiple product entries, often with several variations of product and manufacturer names. Redundant entries were eliminated by designating one version of a product or manufacturer name as the preferred (or standard) entry and

Table I. Data Base

FILE	CONTENT
MANUFACTURER / PRODUCT	PREFERRED & SYNONYMOUS MANUFACTURER & PRODUCT NAMES
BOOKKEEPING FILE	STATUS CODES FOR PRODUCT REQUESTS
PRODUCT INGREDIENT FILE	CHEMICAL INGREDIENTS OF REQUESTED PRODUCTS
CHEMICAL DICTIONARY	STANDARDIZED / VALIDATED CHEMICAL INGREDIENT NAMES

then designating all other forms as synonyms, with proper pointers added to the files to link synonymous records. For manufacturer names and addresses, this process was facilitated by using the Dun and Bradstreet "Reference Book of Manufacturers" and local telephone directories as sources for preferred names and addresses of manufacturing firms. These validation and editing activities result in clean Manufacturer and Trade Name Product files.

For the CPSC project, the names and addresses of manufacturers and their products were identified from published reference sources, including directories, trade journals, and promotional literature. Since all names and addresses are standardized using the Dun and Bradstreet "Reference Book of Manufacturers" prior to input, there are no synonyms for manufacturer or product names. Thus, the manufacturer and product names can be associated in a single file, the Manufacturer-Product File, which requires no further editing prior to ingredient request generation.

INGREDIENT DATA COLLECTION


In order to obtain detailed ingredient data for each of the industrial and consumer trade name products, in most cases it was necessary to go directly to the manufacturer. (In the cases of a few products with relatively standard formulations, such as butyl Cellosolve, it was possible to derive ingredient data from reference sources such as Van Nostrand's "Condensed Chemical Dictionary".) It was determined that a mail survey questionnaire technique would be the most efficient method of requesting ingredient data. A high response rate to our questionnaire surveys was expected because they were being sponsored by Federal agencies.

The procedure designed to accomplish the mail survey was computer-assisted and consisted of three steps: (1) request generation, (2) request mailing, and (3) response receipt.

Because of the large volumes of data to be handled, the generation of a request for each unique product was delegated to the computer. This also allowed for the status of the responses to be automatically monitored. The nucleus of the request generation and monitoring subsystem is a "Bookkeeping" file. Preferred products and manufacturer names and addresses are entered into this file, which then can be employed for request generation and monitoring. Request labels for mailing are then generated by computer for each new entry into the monitoring system. These labels are used in the assembly of product ingredient request packages, illustrated in Figure 2, containing an explanatory letter, one request form for each product (identified by label), and a franked return envelope, which are mailed to the manufacturers.

As responses are received from manufacturers, they are

NATIONAL OCCUPATIONAL HAZARD SURVEY
 c/o Auerbach Associates, Inc.
 121 North Broad Street, Room 928
 Philadelphia, Pennsylvania 19107

 **DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE**
PUBLIC HEALTH SERVICE
CENTER FOR DISEASE CONTROL

NATIONAL INSTITUTE FOR OCCUPATIONAL SAFETY AND HEALTH
 5600 FISHERS LANE
 ROCKVILLE, MARYLAND 20852

PRESIDENT
AUERBACH ASSOCIATES, INC.
 121 N BROAD ST
 PHILADELPHIA, PA 19107

January 30, 1976

2 TRADE NAME REQUEST(S) ENCLOSED

Gentlemen:

For th
 Safety
 Hazar
 Safety
 out th
 used
 brand
 ingred

To ef
 Sectio
 in and
 appro
 that
 we ha
 supply
 form
 Non-s
 when
 manu
 mator
 for m
 ingred
 name

If any
 trade
 and de
 For es
 speci
 secret
 15 of

LABYNE PAPER PREPARATION

REQUEST NUMBER 53041236171234

ASSURANCE OF CONFIDENTIALITY - Any NIOSH contractor having access to this information is legally required by contract to hold all such information confidential.

HAGU DEBUGGER

REQUEST NUMBER 663041231555999
 CONTAINS AN AEROSOL PROPELLANT

ASSURANCE OF CONFIDENTIALITY - Any NIOSH contractor having access to this information is legally required by contract to hold all such information confidential.

	COMPONENT NAME														APRX PER CENT	P	T	NIOSH ONLY			
	15	16	18	20	22	24	26	28	30	32	34	36	38	40				42	43	44	46
1.																					
2.																					
3.																					
4.																					
5.																					
6.																					
7.																					
8.																					
9.																					
10.																					
11.																					
12.																					
13.																					

CDC 2.1
10/73

☐ Patent exists on this product.

☐ This product analysis contains trade secret information. The nature of this information is described on an accompanying sheet.

☐ If you have used supplemental sheets, please indicate how many.

CDC 2.1 (NIOSH)
10/73

Form Approved
OMB No. 68-R-1337

Figure 2. Product ingredient request package.

logged into the Bookkeeping file with an appropriate code to indicate the response status.

MONITORING RESPONSE STATUS

Codes are assigned (either manually or by computer) to

denote the response status for each product request mailed. The following status categories are employed:

- first request (automatically entered when mailing label is generated)
- second (follow-up) request (automatically assigned)

- nonresponse to first and second requests (automatically assigned)
- response containing usable chemical information (manually assigned)
- response requiring further ingredient clarification (manually assigned)
- satisfactorily edited and completed response (automatically assigned)
- unidentifiable product (manually assigned)
- request deleted from system (manually assigned)
- request returned by Post Office (manufacturer cannot be located) (manually assigned)

Each time the monitoring file is updated, default conditions arising from the absence of a manually assigned response code result in the generation of an automatic response code, as well as the manually assigned response codes. Follow-up communications by mail or phone then take place and may result in a change of response status. A Monitoring Statistics Report is also generated following each file update, which gives numerical counts of the number of products in each response category.

CHEMICAL EDIT AND VOCABULARY CONTROL OF INGREDIENT RESPONSES

All responses containing usable chemical information undergo editing by a staff of chemists. The purpose of the chemical edit, whose procedures are illustrated in Figure 3, is threefold: (1) to ensure completeness, accuracy, and specificity of the information provided by manufacturers; (2) to standardize chemical nomenclature for ease of data storage and retrieval; and (3) to ensure that the ingredient information is coded in the proper format for subsequent computer processing.

For the CPSC project, a standardized input vocabulary was not required. In order that an ingredient response be accepted as valid, guidelines for the review and verification of chemical terminology have focused on specificity, accuracy, and completeness. Standardization of chemical nomenclature and synonym association will be provided at the completion of the project by processing verified ingredient names through the Name Match Program developed by Chemical Abstracts Service.

Because of the size of the final NIOSH data base (up to 86 000 products containing an average of six ingredients each), it was considered essential to reduce the number of chemical ingredient synonyms reported by means of a standardized vocabulary. The minimum requirement for this vocabulary was that it provide an authority list relating synonyms to preferred chemical terms. In addition, it was desirable for the vocabulary to list chemical compounds in hierarchical order, based on chemical structure, to allow data retrieval to be either generalized or specialized. Accordingly, a hierarchical thesaurus of approximately 12 000 chemical compounds appearing in the course of the National Occupational Hazard Survey was developed.

To date, the NIOSH project has received over 40 000 responses containing chemical information. Use of the NIOSH thesaurus has reduced the number of different chemical compounds reported from a potential total of 240 000 to an actual total of approximately 8000. The remaining 4000 chemical terms in the thesaurus are synonyms to the 8000 preferred terms.

DATA BASE COMPILATION

Computer software employed on these projects consists of 52 COBOL programs for the NIOSH project and 25 for the

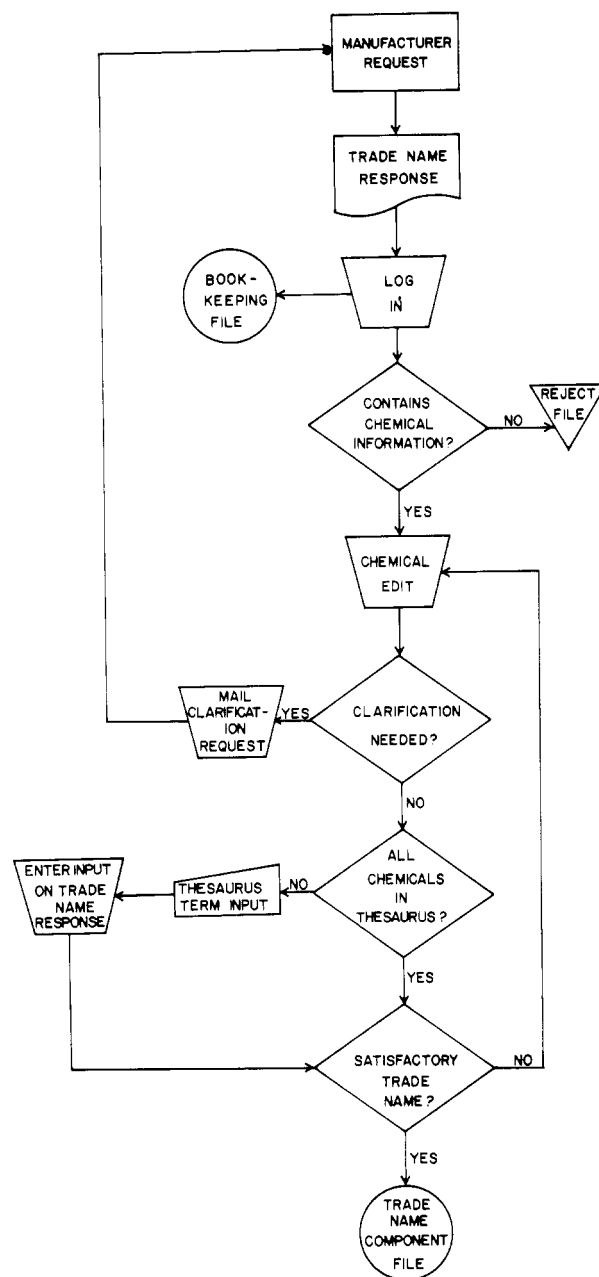


Figure 3. Chemical edit procedures.

CPSC project. These are designed for operation on an IBM 370-168 or 370-165 computer. Twelve of the NIOSH programs relate to the development of the thesaurus. The software design fulfills three objectives: (1) to reduce clerical effort; (2) to reduce redundant data to a unique, preferred format; and (3) to convert acquired chemical ingredient data into a condensed format for ready storage and retrieval.

The data base consists of five major files for NIOSH and four major files for CPSC, as summarized in Table I. These are: a manufacturer file, a trade name product file (or, for CPSC, a combined manufacturer-product file), a bookkeeping file, a product ingredient file, and a chemical dictionary.

The product ingredient file consists of identification numbers of products for which valid responses have been received, chemical code numbers for their respective ingredients, and percentages of these ingredients.

For the NIOSH project, the chemical dictionary file is an extract of all preferred chemical terms in the thesaurus and their identification codes. Only a preferred term is considered a valid entry for a product ingredient, and all chemical code numbers in the trade name component file must exist in the

chemical dictionary.

For the CPSC project, the chemical dictionary consists of an unstructured machine-readable file of acceptable chemical nomenclature for product ingredients. The terms are listed in alphabetical order and compiled in conjunction with the chemical edit. As new chemical terms are derived from the chemical edit, they are added to the chemical dictionary without regard to synonymous or hierarchical relationships, providing they are valid and specific. Upon project completion, the chemical dictionary will be processed through the Name Match Program to standardize nomenclature and associate synonymous terms.

Computerized editing routines, which check the individual files for completeness and compatibility with each other and with the manufacturer and product input data, are an important element of the data base development system.

This work, which has been conducted for both NIOSH and CPSC, in compiling data bases containing the chemical ingredients of over 100 000 trade name products, represents a significant and perhaps unprecedented undertaking. Printed dictionaries of trade name products and ingredients have been published in the past, but these have been neither as exhaustive nor as specialized as the NIOSH and CPSC efforts, and they have not been computer searchable by either product name or ingredients. In contrast, the final result of both the NIOSH and CPSC projects is a machine-readable, standardized, and highly specific data base.

This paper represents the views of Auerbach Associates, Inc., and does not necessarily reflect the views of the government agencies concerned.

Production of a Hierarchical Chemical Thesaurus[†]

HERBERT B. LANDAU* and WENDY L. BYER

Auerbach Associates, Inc., Philadelphia, Pennsylvania 19107

Received March 23, 1976

Through the utilization of computer-aided thesaurus building techniques, a standard vocabulary of chemical nomenclature was established to facilitate indexing and retrieval for a data base which defines worker exposures to organic and inorganic chemical compounds. Developed for the National Institute for Occupational Safety and Health, this vocabulary control tool, in the form of an information retrieval thesaurus, contains 12 000 terms with approximately 8000 preferred descriptors and approximately 4000 chemical synonyms. The thesaurus follows the ANSI Standard as to structure and cross-referencing conventions and was constructed according to rigorous lexicographical procedures employing a methodology built around a computer system known as the Hierarchical Indented Thesaurus System (HITS). The thesaurus features a hierarchical indented term display, a term "tree structure", and automatically generated hierarchical decimal classification codes.

INTRODUCTION

Since July 1973, AUERBACH has been conducting the Trade Name Ingredient Clarification (TNIC) Project for the Hazards Surveillance Branch of the National Institute for Occupational Safety and Health (NIOSH) (Landau¹). The purpose of this study is to identify and record the specific chemical ingredients of up to 86 000 trade name industrial products recorded as exposures to workers during the National Occupational Hazard Survey.

To ensure maximum utility and retrievability of the resulting ingredient data base, component information reported by manufacturers must be standardized. In addition, editing of manufacturers' reports to ensure that valid and specific chemical nomenclature is employed requires some form of vocabulary control. It was therefore decided early in the project to create a chemical thesaurus, with a full hierarchical and cross-reference structure, to serve as the means of converting the diverse chemical nomenclature reported by manufacturers into a single logically structured vocabulary. This chemical thesaurus, known officially as the "Exposure Dictionary of the National Occupational Hazard Survey" (EDNOHS), has grown as the project progressed. It now fills

approximately 1500 pages of computer printout and consists of approximately 8000 main postable (i.e., preferred) terms, 4000 synonyms, and over 55 000 cross references and is employed as a vital working tool in this project.

THESAURUS PURPOSE AND FUNCTION

By common definition, an information retrieval thesaurus is a "...compilation of words and phrases showing synonymous, hierarchical, and other relationships and dependencies, the function of which is to provide a standardized vocabulary for information storage and retrieval".²

Therefore, when the need for some form of vocabulary control tool for trade-name product ingredients became evident early in the NIOSH TNIC project, it appeared that the information retrieval thesaurus concept might be suited to our needs.

A thesaurus could provide a means of resolving the cross-industry (and in many cases intra-industry) synonymy present in the diverse ingredient nomenclature submitted by manufacturers (e.g., Burnt Lime vs. CaO vs. Calcium Oxide; EtOH vs. Ethanol; Melaniline vs. 1,3-Diphenylguanidine; Chlorophenol Sodium Salt vs. Sodium Chlorophenate, etc.). It could also ensure that those who will search the chemical ingredients data base will be speaking the same language as those who built the data base by providing a common standard vocabulary. In addition, since chemical elements and com-

[†] This work was performed under National Institute for Occupational Safety and Health Contract No. HSM-99-73-67. This paper was presented at the 10th Middle Atlantic Regional Meeting of the American Chemical Society, Philadelphia, Pa., Feb 24-25, 1976.