Assessing the Feasibility of Obtaining Product Ingredient Data from Publicly Available Sources†

WENDY L. BYER
ROMAR Consultants, Inc., Technical Information Division, Philadelphia, Pennsylvania 19102

Received May 7, 1982

In fulfillment of its mandate to protect workers from unreasonable risks of exposure to chemicals, a federal research agency has instituted an ongoing study to determine the chemicals used in all types of American industries. Because a large percentage of "chemicals" used are actually formulated trade name products, this study has necessitated identification of their individual ingredients. In previous years, this has been accomplished by requesting the information from product manufacturers. To reduce the burden of these reporting requirements on manufacturers and to expedite the collection of information, the agency has sponsored the work reported herein to investigate the feasibility of acquiring product ingredient data from existing sources.

INTRODUCTION

In fulfillment of its mandate to protect workers from unreasonable risks of exposure to chemicals, the National Institute for Occupational Safety and Health (NIOSH) has instituted an ongoing study to determine the chemicals used in all types of American industries. Because a large percentage of "chemicals" used are actually formulated trade name products, this study has necessitated identification of their individual ingredients. In previous years, this has been accomplished by requesting the information from product manufacturers, with the collected information incorporated into a trade name ingredient data base, the core of which is the Trade Name Component File. In order to eliminate, or at least reduce, the burden of these reporting requirements on manufacturers and to expedite the collection of information, the agency has sponsored work to investigate the feasibility of acquiring product ingredient data from existing sources. Any trade name products whose ingredients could be identified by using publicly available information sources would be considered to have undergone "generic" resolution, as opposed to resolution based on manufacturer-generated reports.

Regardless of the outcome of the feasibility study, NIOSH planned to retain the option of requesting ingredient data directly from manufacturers. This mode of information gathering is necessary for products whose ingredients are not divulged publicly because of their trade-secret status.

The feasibility assessment consisted of the following steps: (1) identification of published, unpublished, and on-line data sources via an exhaustive literature search and personal contacts; (2) selection of a sample of products from the trade name ingredient data base with which to evaluate and compare data sources; (3) development of an assessment methodology to validate the reliability of each source, identify its benefits and/or limitations, and determine overlap of coverage among sources; (4) sequential matching of increasing numbers of sample products to permit early elimination of the least promising data sources and detailed examination of more promising sources. After the most useful sources were identified, they were ranked according to the extent to which they satisfied the assessment criteria, and the data processing requirements and costs associated with using each source were determined.

IDENTIFICATION OF EXISTING DATA SOURCES

Existing compilations of trade name component data were sought through the following sources: government agencies, trade associations, commercial data bank producers and vendors, commercial producers of trade name products, and books and other publications.

Fifteen government agencies and government-sponsored organizations were contacted by telephone to obtain recommendations for trade name ingredient information sources.

Twenty-five general chemical and specialty trade associations were selected to be contacted on the basis of the author's previous knowledge of and experience with such associations during the performance of other government contracts and a review of *Gale's Encyclopedia of Associations*.

On-line searches of commercial data bases were conducted both to identify printed sources of trade name ingredient data and to evaluate the usefulness of selected data bases as direct sources of data.

At NIOSH's request, several large chemical product manufacturing companies were contacted to determine whether they were able and willing to provide portions of in-house trade name data bases or printed information for the purpose of assisting NIOSH in (1) comparing current formulations with formulations reported during a previous study to determine their rate of change, (2) determining formulations of products for which no data was received previously, and (3) verifying manufacturer/trade name correlations reported by surveyors while they are still in the field during the current study. None of the companies contacted agreed to provide any of the requested information.

Printed sources of trade name ingredient information were identified by three routes: (1) on-line data base searches, (2) recommendations of personal contacts at government agencies and trade associations, and (3) a search of *Books in Print*. The subject guide to the 1979–1980 edition of *Books in Print* was searched by using the following terms: branded merchandise; brand names (see trade-marks); chemistry-dictionaries; chemistry, technical-dictionaries; chemistry, technical-formulae, receipts, prescriptions; product safety; trade-marks. (No appropriate headings were found for "component", "composition", "formula", "formulation", or "ingredient".)

Results of the data source identification effort were as follows.

The majority of potential trade name ingredient sources were identified through telephone interviews with representatives of government agencies and trade associations. Although computerized and manual literature searches were also performed, these yielded very few new sources.

The individuals contacted were librarians, information specialists, and technical personnel. They were asked about the existence of data sources produced by their organizations, as well as any other potential sources of which they were aware. Nearly everyone contacted was able to offer suggestions of possible sources, except for those at some of the trade

† This work was performed under National Institute for Occupational Safety and Health Contract No. 210-80-0059. This paper was presented at the 16th Middle Atlantic Regional Meeting of the American Chemical Society, Newark, DE, April 21–23, 1982.

0095-2338/82/1622-0190$01.25/0

Table I. Potential Trade Name Ingredient Data Sources

Printed Sources	
ACGIH Trade Names Index	Kirk-Othmer Encyclopedia of Chemical Technology
AMA Drug Evaluation	Materials Handbook
American Drug Index	The Merck Index
The Chemical Formulary	Modern Plastics Encyclopedia
Chemical Synonyms and Trade Names	National Formulary
Chemical Trade Names and Commercial Synonyms	A New Dictionary of Chemistry
Chemical Week 1980 Buyers' Guide	NIOSH/OSHA Pocket Guide to Chemical Hazards
Chem Sources	1979-80 OPD Chemical Buyers' Directory
Colour Index	Physicians' Desk Reference
The Condensed Chemical Dictionary	PLASTEC Note N9C. Trade Designations of Plastics and Related Materials
Cosmetic Ingredient Dictionary	
CRC Handbook of Food Additives	POISINDEX
Dangerous Properties of Industrial Materials	Registry of Toxic Effects of Chemical Substances
Drug Topics Red Book	SOCMA Handbook of Commercial Organic Chemical Names
Encyclopedia of International Health and Safety	Thomas Register of American Manufacturers and Thomas Register Catalog File
FDA Approved Drug Products	
Fenaroli's Handbook of Flavor Ingredients	Thorpe's Dictionary of Applied Chemistry
General Tire and Rubber Co. Trade Names Collection	Trade Names Dictionary
Hackh's Chemical Dictionary	The United States Dispensatory
Handbook of Material Trade Names	The United States Pharmacopeia
Handbook of Nonprescription Drugs	
Data Bases	
BIOSIS Previews	Chemname
CA Search	Claims/Chem
Chemdex	Clinical Toxicology of Commercial Products
Chemical Industry Notes	Enviroline
Chemline	PTS Promt

associations which were only membership headquarters with no technical staff.

Chemical and trade name dictionaries, handbooks, and encyclopedias were by far the most commonly recommended sources. Very few government-sponsored sources were identified. The Consumer Product Safety Commission's product ingredient data base was never suggested, nor was NIOSH's own trade name ingredient data base. Government spokesmen, in particular, mentioned EPA's Toxic Substances Control Act (TSCA) Inventory of Chemical Substances. However, the Inventory, while including trade names, contains no ingredient information. Poison information centers and the POISINDEX, a microfiche collection of trade name product ingredients for hospital emergency room reference, were commonly mentioned by government contacts. Only a few people suggested the *Clinical Toxicology of Commercial Products*, and they seemed unaware of the government-sponsored computerized version.

Trade associations relied heavily on both general and specialized buyers' guides. None of the associations contacted produced its own trade name ingredient compilation, and none maintained a list of member companies' trade names, referring instead to the buyers' guides.

One government-sponsored information source, suggested by a trade association, was the Plastics Technical Evaluation Center (PLASTEC). This organization produces a book listing plastics trade names, a brief description of the item, and the manufacturer. Also identified was an unofficial source in private industry. The librarian at the General Tire and Rubber Co. has been collecting trade name information for a number of years and welcomes requests. Although this collection concentrates on trade names related to the rubber industry, products from other areas have also been included.

In all, 39 printed sources and 10 data bases were identified as potential sources of trade name ingredient information. These are listed in Table I.

METHODOLOGY FOR ASSESSING DATA SOURCE FEASIBILITY

Having identified potential sources of trade name ingredient information, it was necessary to assess the feasibility of using them as input to NIOSH's Trade Name Component File. The methodology developed for feasibility assessment was a three-step process. The first step involved selecting a representative sample of trade names from the existing Trade Name Component File for matching against potential ingredient sources; the second step comprised matching the sample; the third step involved evaluating the results of the match (for each source having matching trade names) against a number of criteria which a source must satisfy in order to be considered feasible.

Trade Name Sample Selection. The purpose of using a sample from the NIOSH data base was (1) to estimate the extent of coverage of NIOSH-identified trade names provided by potential trade name ingredient data sources and (2) to test the reliability of a source against manufacturer-supplied ingredient data in the NIOSH data base.

The sample was selected by means of specially written programs as follows: (1) Secondary trade names (i.e., a trade name product used as an ingredient in a NIOSH-identified trade name) were eliminated as being unrepresentative of field-identified trade names. (2) Primary (i.e., field-identified) trade names containing unresolved secondary trade names were eliminated. (The remaining trade names were those that had been fully resolved; i.e., their chemical ingredients had been identified.) (3) Of the remaining fully resolved trade names, one out of every 100 single-component trade names and one out of every 300 multicomponent trade names were selected and printed in two separate lists. (4) The lists were merged into a single list of 180 trade names, each of which was numbered 1, 2, 3, or 4 in sequence, creating four subsets of 45 trade names each. (5) After sample trade names were matched against data sources, as described in the next section, NIOSH printed a list of ingredients for those NIOSH trade names which were also listed in the data sources. Trade name products whose ingredients were denoted as confidential by manufacturers reporting to NIOSH were not included in this list but were checked in-house by NIOSH for extent of ingredient matching with the data sources.

Trade Name Matching Procedure. The matching operation was conducted in a stepwise manner by using the list of 180 representative trade names. The sample was divided into sets of approximately 45 trade names each and used sequentially in the matching operation.

Table II. Data Sources Used for Step 2 Matching

Printed Sources
ACGIH Trade Names Index
AMA Drug Evaluation
American Drug Index
Chemical Synonyms and Trade Names
Chemical Trade Names and Commercial Synonyms
Chemical Week 1980 Buyers' Guide
Colour Index
The Condensed Chemical Dictionary
Drug Topics Red Book
FDA Approved Drug Products
General Tire and Rubber Co. Trade Names Collection
Handbook of Material Trade Names
Kirk-Othmer Encyclopedia of Chemical Technology
Materials Handbook
Modern Plastics Encyclopedia
Physicians' Desk Reference
PLASTEC Note N9C. Trade Designations of Plastics and Related Materials
POISINDEX
Registry of Toxic Effects of Chemical Substances
SOCMA Handbook of Commercial Organic Chemical Names
Trade Names Dictionary
The United States Dispensatory

Data Bases	
BIOSIS Previews	Chemname
Chemical Industry Notes	Enviroline
Chemline	PTS Promt

Table III. Trade Name Ingredient Data Sources Selected for In-Depth Feasibility Assessment

American Drug Index
Chemical Synonyms and Trade Names
Chemname (On-line data base)
Chemical Industry Notes (On-line data base)
Colour Index
The Condensed Chemical Dictionary
FDA Approved Drug Products
PLASTEC Note N9C. Trade Designations of Plastics and Related Materials
POISINDEX

In step 1 of the match, one set of 45 trade names was matched against each of the 49 potential data sources. If no more than one match occurred, the data source was considered to be unusable for trade name ingredient resolution, and no further evaluation took place for the source; 21 data sources were eliminated in this manner.

In step 2, a second set of 45 trade names was matched against the remaining 28 data sources, as shown in Table II.

In step 3, the most promising data sources were selected for an expanded match and in-depth feasibility assessment. The nine sources selected are shown in Table III. Promising sources were selected primarily on the basis of the number of sample trade names having hits, either for general trade name product types (e.g., Condensed Chemical Dictionary) or for trade name products specific to an industry (e.g., American Drug Index).

FEASIBILITY ASSESSMENT PROCEDURE AND RESULTS

Following the matching operation, data sources were evaluated against a set of criteria to determine the feasibility of using them. Evaluation criteria included the following: extent of matching between the NIOSH Trade Name Component File and the data source, benefits and/or limitations of the source, reliability of the source, effective life of the data provided by the source, the extent of overlap between sources (see Table IV), and data processing requirements and costs.

Extent of Product Matching. This was an important consideration because ingredients of particular trade names were being sought. Table V shows the extent of product matching in data sources on using the complete 180 product sample. Data sources are listed in descending order, according to number of trade names having a match in the source. The number of personnel hours expended per match is also given.

Table IV. Relationship between Number of Trade Name Hits and Amount of Overlap

no. of hits	highest incidence of overlap, %	overlapping sources
2	100	ADI/FDA and Chemname/FDA
3	33.3	Chemname/CI
11	36.4	Chemname/CIN
12	16.7	ADI/Chemname
16	62.5	CCD/CSTN
17	64.7	CCD/Plastec
19	36.8	Chemname/CCD and CIN/CCD

Table V. Extent of Trade Name Product Matching Using a Sample Size of 180

name of source	no. of matches	% of sample having matches	personnel hours per match
Condensed Chemical Dictionary	21	11.7	0.14
PLASTEC	20	11.1	0.12
Chemical Synonyms and Trade Names	19	10.6	0.16
POISINDEX	17	9.4	0.24
Chemname	15	8.3	0.23
Chemical Industry Notes	12	6.7	0.65
Colour Index	4	2.2	0.69
American Drug Index	3	1.7	0.22
FDA Approved Drug Products	2	1.1	0.38

Benefits and/or Limitations of Sources. An analysis of the benefits and limitations (including any restrictions on use) was performed for the 28 ingredient data sources which had at least one match. Information regarding restrictions on use was obtained in response to a form letter sent to publishers of printed sources and producers of automated data bases.

In general, the most productive and easy to use data sources were printed sources in dictionary or abbreviated encyclopedia format. Although few matches occurred for this type of source when the source was specific to certain industry sectors (e.g., American Drug Index, Colour Index), the lack of matches was attributed to their specialization. If NIOSH trade names could be identified as drugs, dyes, plastics, etc., these sources could be productively used for selective searching.

For the most part, the automated data bases were neither convenient to use nor productive. After on-line searches of bibliographic data bases, an original article must usually be located and scanned, because the data base often gives only citations. Two exceptions were the Chemname and Chemical Industry Notes data bases. The former provides immediate identification of the ingredient of single-ingredient trade names, and the latter often has abstracts containing ingredient information.

Limitations encountered to the use of some data sources fall into the following categories: (1) lack of ingredient information for trade name products (e.g., Chemical Week Buyers' Guide, Modern Plastics Encyclopedia); (2) obsolescence of trade name ingredient data (e.g., Chemical Trade Names and Commercial Synonyms, Trade Names Index); (3) need to search multiple volumes (e.g., Kirk-Othmer Encyclopedia of Chemical Technology); (4) need to locate and review original journal articles after a match has occurred (e.g., CA Search, Enviroline).

Table VI. Scale for Scoring Reliability

characteristic	score
source contains 100% same ingredients (by item) as NIOSH data	1
source contains 75–99% same ingredients as NIOSH data	2
source contains 50–74% same ingredients as NIOSH data	3
source contains 25–49% same ingredients as NIOSH data	4
source contains less than 25% same ingredients as NIOSH data	5

Few restrictions on the use of copyrighted material were encountered. Many of the data base producers charge royalties for information extracted from their data files if the information is to be further disseminated. A number of book publishers would have to know what the information was to be used for and give written permission for such use. However, most would not refuse permission.

Reliability of Sources. A reliable source is one which is both accurate and sufficient in its ingredient reporting. That is, not only must the reported ingredients be correct but also all, or nearly all, of a product's ingredients must be listed.

As a test of reliability, the ingredients reported for sample trade names having a match in one or more data sources were compared with the ingredients reported to NIOSH for that product by the manufacturer. Those sources whose reported ingredients most closely matched the NIOSH ingredient records were considered the most reliable. (Products designated as confidential by manufacturers were evaluated and scored by NIOSH.)

Reliability was measured on an objective scale, shown in Table VI. For example, if a product had four ingredients reported in the Trade Name Component File but only three of these (identical) ingredients were reported in a data source, the degree of matching would be 75%, and the data source would receive a score of 2.

In the scale, the characteristics of a source are arranged in descending order of desirability and are numbered sequentially. Thus, the lower the score of a source, the better is its reliability. To determine the overall reliability of a source, the reliability of each matching trade name in the source was scored, and then the scores were averaged across all trade names. Sources were then ranked in order of reliability.

The ranked results of reliability scoring are presented in Table VII. (Although 28 sources were scored and ranked, only the nine sources selected for in-depth study are presented here for purposes of illustration.)

The reliability measurement effort indicates that the sources with the greatest degree of reliability are those which are either specific to an industrial sector (e.g., American Drug Index, Colour Index) and/or identify the ingredient of single-ingredient trade names, rather than formulated products (e.g., Colour Index, Chemname).

Effective Life of Data Provided. The most important indicator of data source reliability, next to correlation of product ingredients between the source and the Trade Name Component File, is the currency of ingredient information. The best method of determining currency is to ask the trade name product manufacturer to verify the ingredients reported in the data source. On the basis of past NIOSH experience with manufacturers, this method was expected to require more elapsed time than was allocated in the project schedule and to be minimally productive, due to the inability of many manufacturers to respond to any additional government reporting requests. The next best approach was to ask data source publishers how frequently ingredient information is revised. To this end, letters were sent to publishers of the 28 data sources in which product ingredient matches were found.

Of primary interest is the finding that the data sources with the highest reliability scores are also among the most up-to-date sources. The converse is not true, i.e., currency of ingredient information does not confer reliability, at least not by the reliability measurement method used in this project.

Overlap between Sources. The amount of overlap between data sources was considered in order to determine whether a potential for duplication of effort exists, should NIOSH decide to use more than one source in its in-house ingredient identification process.

Table VIII shows the amount of overlap among the nine data sources selected for expanded feasibility assessment. Unfortunately, there is no clearcut correlation between number of trade name hits and the likelihood of overlap, as indicated in Table II.

Because the greatest amount of overlap is barely 65%, it would seem risky to eliminate any of these potential data sources on the basis of percent overlap alone. (The case of 100% overlap based on two hits is too small a sample to be reliable for American Drug Index/FDA and Chemname/

Table VII. Reliability Ranking of Trade Name Ingredient Sources for Sample Size of 180

source	trade name[a] (score)									av score
	TRN 1	TRN 2	TRN 3	TRN 4	TRN 5	TRN 6	TRN 7	TRN 8	TRN 9	
American Drug Index	Betadine (1)	Liver Ex. (1)								1
Colour Index	D&C (1)	Murexide (1)								1
FDA Approved Drug Products	Betadine (1)									1
Chemname	Betadine (1)	Cymel (5)	Marlex (1)	Murexide (1)	Polyhall (5)	Tergitol (1)	Unads (2)	Wattle (5)		2.63
Chemical Synonyms and Trade Names	Betadine (1)	Duraplex (1)	Fr. Ochre (4)	Grinding (5)	Loctite (5)	Marlex (1)	Murexide (5)	Neville (1)	Tergitol (5)	3.11
Condensed Chemical Dictionary	Cymel (5)	Imron (5)	Duraplex (1)	Herchlor (1)	Loctite (5)	Marlex (1)	Staclipse (5)	Tergitol (5)	Unads (2)	3.33
PLASTEC	Coverlac (5)	Cymel (5)	Imron (5)	Duraplex (1)	Herchlor (1)	Instaweld (1)	Mark (5)	Marlex (1)	Petro (x)	
	Scotchseal (2)	Slide (5)	Tergitol (5)							3.55
POISINDEX	Betadine (1)	Linseed (5)	Dietzgen (5)	Imron (5)	Lt. Min. oil (1)	Pzazz (5)	Slide (4)	Squibb Min. oil (1)	West Soap (5)	3.56
Chemical Industry Notes	Imron (5)	Duraplex (5)	Herchlor (1)	Instaweld (1)	Marlex (1)	Scotchseal (5)	Tergitol (5)	Unads (5)	Wattle (5)	3.67

[a] Abbreviated TRN.

Table VIII. Overlap between Nine Data Sources Based on Sample Size of 180 Trade Names

source no.	source [no. of hits]	source no. [no. of hits]								
		1 [3]	2 [12]	3 [11]	4 [3]	5 [19]	6 [2]	7 [16]	8 [17]	9 [16]
1	American Drug Index (ADI) [3]		2	0	0	1	2	2	0	2
2	Chemname [12]			4	1	7	2	6	5	2
3	Chemical Industry Notes (CIN) [11]				0	7	0	4	8	1
4	Colour Index (CI) [3]					0	0	1	0	0
5	Condensed Chemical Dictionary (CCD) [19]						1	10	11	2
6	FDA Approved Drug Products [2]							2	0	2
7	Chemical Synonyms and Trade Names (CSTN) [16]								6	2
8	PLASTEC [17]									2
9	POISINDEX [16]									

FDA. However, it is highly likely that the latter source and the American Drug Index would duplicate each other extensively, allowing one of these sources to be dropped.) Of course, the costs of using these sources for in-house ingredient resolution significantly affects the implementation decision.

Data Processing Requirements and Costs. Two types of trade name ingredient data sources were evaluated: (1) automated (on-line or magnetic storage) and (2) hard copy (book format or microfiche). The single microfiche source, POISINDEX, can also be provided as an on-line data base.

Automated Data Processing Requirements. To access information in on-line sources, the data base would have to be queried for each trade name and synonymous trade name, necessitating entry of trade names via terminal (or corresponding mechanism) either individually or batched on a magnetic device such as a diskette. Responses could be captured on a magnetic medium such as a diskette and then processed by means of a computer program. Alternatively, responses would be output in hard copy form and would subsequently have to be transcribed to computer-readable format.

While on-line searching of automated data bases is satisfactory for a small sample of trade names, it would probably not be the most cost-effective method for matching the 40 000–100 000 trade names likely to be identified each time NIOSH updates its industry-wide survey of chemical exposures. For such large-scale searches, the most efficient method would be a batch-mode search, conducted for NIOSH by a data base producer or vendor or conducted by NIOSH using magnetic tapes leased or purchased from a data base producer. For this type of search, NIOSH would produce a magnetic tape extract of that portion of its data base containing trade names and selected synonymous names for matching against the data base tapes.

If a source is available on magnetic storage, either NIOSH or the data base producer or vendor would have to write a program to extract those trade names of interest, develop any synonymous trade names (if desired), and match them against the data base. Either party would then run the program. Other software would have to be developed by NIOSH to process matches.

Information about the cost of purchasing or leasing data base files was solicited from data base producers and vendors, who were also asked whether they would conduct batch searches for trade names in-house and what such a service would cost. This information was used to generate cost estimates for searching selected data bases.

Hard Copy Data Source Processing Requirements. Since it is impractical to look up thousands of trade names in one or more books, the most appropriate means of handling trade name searches in books is to convert all or part of the information contained in a book to machine-readable format and then run a matching program. There are two possible levels of format conversion.

(*1*) *Conversion of Trade Names Only.* In this case, any matches will require a lookup of ingredients in the book. Ingredients would then have to be entered into a computer file via transcription. While the use of optical scanning equipment for this purpose is a possibility, the state-of-the-art would probably make this an expensive and/or impractical method.

(*2*) *Conversion of Both Trade Names and Ingredient Data.* In this case, a match would automatically generate a list of trade name ingredients, which then would be entered into a computer file, either automatically or manually. However, the cost of converting the ingredients of all trade names in a source to machine-readable form would be quite high, with no guarantee of a sufficiently high proportion of matches.

If a source is available on microfiche only, the processing requirements would be much the same as those for book format sources. In many instances, however, microfiche sources are also available on magnetic storage, because the microfiche has been generated from a computer file. This is true of POISINDEX. In such cases, the data source can be processed in the same way as other magnetic storage sources.

Cost estimates were developed for processing hard copy sources in two modes: (1) automated matching of trade names with manual transcription of matching trade name ingredient information and (2) manual matching of trade names with manual transcription of matching trade name ingredient information. (Transcription time requirements are identical in either mode.)

Estimates for automating data sources took into account the cost of the source, the cost to convert the source to machine-readable format, the cost to write a program to match source terms with NIOSH trade names, the cost for clerical personnel to look up the ingredients of matching trade names in a printed source, and the cost to perform manual transcription of ingredients onto a trade name ingredient reporting form or to keyboard ingredients on a terminal.

Estimates for performing manual matching were based on the actual time spent in looking up the 180 sample trade names in the nine sources selected for in-depth evaluation. (Note that these times may vary for larger batches of trade names, due to the fatigue element.) The cost of a source was omitted in these estimates, because this cost is miniscule compared to the cost of labor.

In every case, automated matching was estimated to be substantially cheaper than manual matching, with savings ranging from $9500 to over $32 000 per source over the latter method. Although automation requires a greater initial dollar investment, it also saves significant amounts of personnel time and thus alleviates many of the management problems associated with labor-intensive activities.

CONCLUSIONS

As a result of the foregoing feasibility assessment procedure, the following conclusions were reached.

The costs of using many of the data sources are substantial, but it is cheaper in every case to automate a source. While requiring an initial investment, automation need only be done once in order for the data to be permanently available. (Of

course, some data base files have an annual lease cost, and the printed sources are revised from time to time.)

The cost per match for some sources may actually be greater than the cost of contacting a manufacturer directly. However, it allows NIOSH to obtain ingredient data much sooner.

Costs of trade name matching using specialized sources could possibly be reduced if products could be categorized according to use. For example, if dyes, pigments, inks, paints, etc. could be classified as coloring agents, only these products could be searched in the Colour Index, eliminating the need to search for all 40 000–100 000 trade names in this source. Also, many additional specialized data sources could be used in generic resolution if the number of trade names to be matched could be limited in this way.

The cost per match for the aggregate data sources may actually be somewhat higher than estimated, due to overlapping coverage among sources. On the other hand, overlap could be beneficial, in that it allows reported product ingredients to be compared and either verified or called into question, in cases of discrepancies.

None of the sources evaluated were found to be totally reliable when compared with NIOSH ingredient data. Practically none provided ingredient percentage information. Depending on the level of accuracy required by NIOSH, manufacturers may eventually have to be contacted directly. However, ingredient data from these secondary sources could still prove useful for interim chemical exposure estimates.

Since most data sources, including the generalized sources, have limited coverage, a combination of sources would have to be used to maximize the number of matches with NIOSH trade names.

All data sources are less current than manufacturer data. (Numerous manufacturers have stated that their product formulations change frequently.)

Use of additional data sources would allow NIOSH to expand its data base to include trade name products not identified during the industrial survey.

The feasibility assessment methodology used in this study might well be of interest to individual firms, public health personnel, and/or labor unions concerned with exposure monitoring of workers. In addition to identifying the ingredients of specific trade name products, the methodology described herein can be adapted to identification of the probable ingredients of certain types of products within various generic use categories such as adhesive and sealant compounds, paints and coatings, and detergents.

User-Oriented Approach to a Computerized Organic Reaction Catalog

BERI J. COHEN

Technicon Instruments Corporation, Tarrytown, New York 10591

Received December 3, 1981

A computerized system for creating and searching an organic reaction catalog is described. Starting materials and products are represented as sets of parameters defining their structural features, written in a notation form. The system deals efficiently with stereochemistry and has extensive capabilities for substructure searching, including searching for closely related functional groups. Its application to small computers is discussed.

INTRODUCTION

The systematic documentation of organic reactions for a computerized retrieval of synthetic information presents a major challenge for organic chemists and information scientists alike.[1-5] The objective is to be able to represent efficiently chemical structures in a retrievable form, so that substructures will be retrieved as well, upon request, in a meaningful way. Two general approaches were described in the literature. One uses key words of partial structural features of the reactants. It is exemplified by the commercially available CRDS system from Derwent Publications[6] and the GREMAS system from IDC.[7] In the first, key words are either trivial names used in organic synthesis jargon or codes of structural fragments according to the "Ringcode".[8] Their limitations as to substructure search in case of complex structures were already discussed.[4,6] The GREMAS system also uses special coding and has considerable versatility in substructure searching. Unfortunately, its high cost makes it unaffordable to most potential users. The other approach is to represent an atom-by-atom structure of the reactant molecules by means of a topological description such as graphs or connection tables. It requires more sophisticated programming and high-performance computers. In recent years many improvements were introduced such as automatic detection of reaction sites,[9,10] use of screens for a preliminary search,[11] and automatic conversion of WLN formulae to connection tables.[12] An in-house system was recently reported from a major company which uses the key word approach,[13] showing that there is no unique, satisfying answer to the problem.

Today, most organic chemists are expected to have access to a computer service. The present work was aimed, therefore, at investigating the possibility of using the key word approach for creating and searching an organic reaction data base in the particular areas of interest of one or several research groups, updating and searching being done by the users themselves. In designing the system, several objectives were set: (a) representation of molecular structures that eliminates the use of dictionaries with a minimum number of rules; (b) easy coding of reaction data and interpretation of computer output after a search has been performed; (c) flexibility and versatility in searches of substructures. We believe that these goals are successfully met by the system described below.

STRUCTURE OF THE CATALOG

Each reaction in the catalog consists of a set of parameters describing the relevant structural features of the starting material, a similar set for the product, a parameter describing the general type of the reaction, and a short text providing additional essential data such as reaction conditions and references. The text can be reduced to a single number referencing a physical storage system such as a card file or microfilm. Apart from this text, all the other parameters are

0095-2338/82/1622-0195$01.25/0