of microfiche. The usefulness of a microfiche of the entire inverted file as a potential user aid could be considered.

Following this same line of thinking, the analysis of terms in the inverted file can be broken down into individual components, for example, by identifying the high-frequency journal title (CODEN's) that appear in a file, or the high-frequency *author's names*, and the distribution of Chemical Abstracts Service (CAS) Registry Numbers that one finds in the inverted file. In each instance, a certain amount of implicit information is available which could influence the user's search strategy. One could easily produce the 100 most highly posted authors in the index file as a guide. In a similar fashion, the 100 most highly posted journal titles (realizing some inherent variability) could also be assembled as a user guide.

Finally, it seems possible that the analysis of term distribution could serve as a valuable tool in weighing the effects of merging one inverted file with another. In the existing software system, used by NLM, the merging of two free-text bibliographic files would essentially consolidate the inverted files, since many of the terms would be common to both files.

LITERATURE CITED

(1) A. D. Booth, "A "Law" of Occurrences for Words of Low Frequency," *Info. Control*, **10**, 4 (April 1967).
(2) Defense Documentation Center, "DDC Retrieval and Indexing Terminology", Vol. I, 1st ed, AD/A-001 200/5GA, 1975.
(3) Defense Documentation Center, "DDC Retrieval and Indexing Terminology", Vol. II, 1st ed, AD/A-001 201/3GA, 1975.
(4) R. A. Fairthorne, "Empirical Hyperbolic Distributions (Bradford-Zipf-Mandelbrot) for Bibliometric Description and Prediction", *J. Doc.*, **25**, 4 (Dec 1969).
(5) H. M. Kissman and D. J. Hummel, "TOXICON - An On-Line Toxicology Information Service", *Chem. Technol.*, 72 (Dec 1972).
(6) P. H. Klingbiel, "Multimillion Word Data Bases: A Preliminary Report", Vol. 1, Defense Documentation Center, AD-777 200/7, 1974.
(7) P. H. Klingbiel, "Multimillion Word Data Bases: A Preliminary Report", Vol. 2, Defense Documentation Center, AD-777 210/6, 1974.
(8) B. Mandelbrot, "An Information Theory of the Statistical Structure of Language", in "Proceedings of the Symposium on Applications of Communication Theory", Butterworth, London, 1953, pp 486-500.
(9) National Technical Information Service, "NTIS Master Frequency List of Subject Terms, July 69-Dec 72", NTIS/SR-74/04.
(10) "On-Line Services Reference Manual", National Library of Medicine, March 1975.
(11) P. B. Schipma, "Term Fragment Analysis for Inversion of Large Files", IIT Research Institute, Chicago, Ill., 1971, 17 pp.
(12) B. M. Vasta, "Use of TOXLINE/CHEMLINE for Retrieving Drug Information", *Drug Inf. Assoc. J.* (1975); presented Oct 28, 1974, at Drug Information Association.
(13) B. M. Vasta, "TOXLINE - NLM's On-Line Information Resource", Proceedings of the National Aeronautics and Space Administration's Annual Conference of Clinic Directors, Environmental Health Officials and Medical Program Advisors, March 18, 1975.
(14) M. E. Williams, "Handling of Varied Data Bases in an Information Center Environment", IIT Research Institute, Chicago, Ill., 1971, 24 pp.
(15) G. K. Zipf, "Human Behavior and the Principle of Least Effort", Addison-Wesley, Cambridge, Mass. 1949.

A Comparison of On-Line and Manual Modes in Searching Chemical Abstracts for Specific Compounds

JOSEPH SANTODONATO

Center for Chemical Hazard Assessment, Life and Material Sciences Division, Syracuse Research Corporation, Syracuse, New York 13210

Received March 10, 1976

Manual searching of *Chemical Abstracts* was compared with computer-searching *CA Condensates* and CBAC, especially with regard to retrieval of broad information on a specific compound. Differences are apparent owing to deficiencies in indexing and in the capability for free-text searching of abstracts in the computer-based systems. The success of manually searching the *Chemical Abstracts* Substance Indexes could not be equalled by the on-line systems, either alone or in combination.

Recent developments in the computerized searching of bibliographic sources which allow for on-line interaction and "browsing" of data bases have aided immeasurably in the efficient retrieval of chemical information. However, inherent limitations in the usefulness of these research tools have become apparent in the searching of specific chemicals as opposed to more generalized search strategies.

While developing a state-of-the-art review on the pesticide toxaphene, we felt that it would be helpful to compare the advantages of manually searching *Chemical Abstracts* with the combined results of machine-searching *Chemical Abstracts Condensates* and the CBAC portion of TOXLINE. To retrieve as many references as possible, we did not restrict our search strategy to provide only the most pertinent citations. In this sense, our literature search was intended to be a quantitative reflection of the published literature relating to toxaphene without making a value judgment regarding the relevance of individual articles.

It has been noted[1,2] that the indexing of *CA Condensates* is less systematic and exhaustive than that of the *Chemical Abstracts* Volume or Cumulative Indexes. In particular, *CA Condensates* appears to be inferior in the retrieval of citations for specific compounds.[3] Our experience has shown this to be especially true when the scope of a search on individual chemicals is extremely broad, requiring information in biological, chemical, and environmental areas. In other comparisons of results with *CA Condensates* and the weekly Keyword Indexes of *Chemical Abstracts*,[4] it was found that in certain instances either method may be superior or both may be equivalent. In quantitative comparisons of our own search results on individual chemicals, we have found that manual searching of Keyword Indexes yields the same number of citations as computerized retrieval because *CA Condensates* is constructed by using only the weekly Keyword Indexes. Superiority of manual searching becomes evident once an annual Volume Index is available. It is important to recognize that the weekly Keyword Indexes of *Chemical Abstracts* are constructed from characterizing words or phrases (keywords) selected from the title or context of an abstract. The Volume Indexes, on the other hand, are derived usually from a searching examination of the original documents, not the abstracts.[5]

The CBAC (Chemical-Biological Activities) data base, which was also used for comparison in the toxaphene search example, is offered through the National Library of Medicine's TOXLINE system. Coverage encompasses Sections 1 through

Table I. Retrieval of Citations for Toxaphene by Various Search Routines

Chem. Abstr. Vol. and year	Citations found manually	Citations found in CBAC	Citations found in *CA Cond.*
76 (1972)	19	10	9
77 (1972)	25	17	7
78 (1973)	35	18	10
79 (1973)	16	9	7
80 (1974)	15	9	7
81 (1974)	25	12	10
82 (1975)	19	15	10
83 (1975)	11	10	11
Total	165	100	71
% of manual citations		61	43

Table II. Distribution of Toxaphene Citations in Various Search Modes

Chem. Abstr. Vol. and year	Citations in CBAC, not in *CA Cond.*	Citations in *CA Cond.*, not in CBAC	Common *CA Cond.*/CBAC citations	Unique citations by manual search
76 (1972)	4	3	6	7
77 (1972)	11	1	6	7
78 (1973)	10	2	8	15
79 (1973)	3	1	6	6
80 (1974)	3	1	6	6
81 (1974)	5	3	5	12
82 (1975)	8	3	6	4
83 (1975)	2	3	8	0
Total	46	17	51	57

5 of *Chemical Abstracts*. CBAC presents distinct advantages in providing the capability for free-text searching of individual abstracts and includes CAS Registry Numbers, but is limited in its coverage of the chemical literature.

SEARCH RESULTS

The keyword toxaphene was searched in TOXLINE and *CA Condensates* (via the Lockheed DIALOG system). A manual search on toxaphene was conducted using the corresponding Annual Substance Indexes of *Chemical Abstracts* for Vol. 76 through 82. Volume 83 (through issue No. 20) was searched by weekly Keyword Index. An analysis of the combined results revealed several interesting relationships among the three search modes.

As indicated in Table I, manual searching provided a considerably greater number of total citations than either mode of computerized retrieval. It is surprising to note that CBAC produced nearly 150% more references than *CA Condensates*, in spite of the fact that CBAC coverage is limited to only five of the 80 sections of *Chemical Abstracts*, all of which are covered in *CA Condensates*. This difference apparently reflects the greater access to articles which is provided by the free-text searching capability in CBAC.

Computer searching for citations on toxaphene in CBAC and *CA Condensates* yielded a number of articles which appeared in only one of the two on-line modes. In addition, several articles were located by machine search which could not be found by manually searching the same keyword. As expected, manual searching alone produced numerous references which were not retrieved from either of the computer-based systems. These results, as summarized in Tables II and III, demonstrate that CBAC provided a much greater number of citations which were not located by *CA Condensates* than vice versa. Manual searching, however, produced 57 references that were missed by both computerized systems combined. Note that for Vol. 83 of *Chemical Abstracts*, no unique citations were found by manual searching, apparently because weekly Keyword Indexes had to be employed.

Table III. Toxaphene Citations Found by Computer but not by Manual Search

Chem. Abstr. Vol. and year	Citations in CBAC	Citations in *CA Cond.*
76 (1972)	0	1
77 (1972)	0	0
78 (1973)	0	0
79 (1973)	0	0
80 (1974)	1	0
81 (1974)	2	2
82 (1975)	1	1
83 (1975)	1	0
Total	5	4

Table IV. Computer Search of Polychlorocamphene Compared to Manual Search of Toxaphene

	CBAC		*CA Condensates*	
Chem. Abstr. Vol. and year	Total no. of citations	Citations also found manually	Total no. of citations	Citations also found manually
76 (1972)	0	0	1	0
77 (1972)	0	0	3	2
78 (1973)	5	3	9	6
79 (1973)	0	0	0	0
80 (1974)	2	2	2	2
81 (1974)	4	0	4	0
82 (1975)	5	0	5	1
83 (1975)	4	0	2	0
Total	20	5	26	11

An attempt was made to increase the computerized retrieval of references for toxaphene by adding the synonymous keyword polychlorocamphene to the machine search strategy. As shown in Table IV, a number of additional citations were identified by searching the new term, many of which had also been located by manual search using only toxaphene as a keyword. It is clear, however, that the combined effectiveness of computer searching two specific chemical synonyms did not equal the success of a manual search employing a single search term.

DISCUSSION

For the investigator whose primary objective is an exhaustive search of the literature relevant to a specific chemical substance, it appears that computer-readable data bases are an inferior substitute for manual searching of *Chemical Abstracts* using the Annual Substance Indexes. In cases where state-of-the-art reviews, criteria documents, etc., are being prepared (i.e., where broad information on specific chemicals is required), the more comprehensive indexing of *Chemical Abstracts* which is available only to the manual searcher offers a decided advantage in retrieving important articles.

It is also evident, however, from the examples presented above, that the free-text abstract searching capability provided in CBAC allows for considerably improved efficiency in computerized citation retrieval when compared to *CA Condensates*. Since the five sections of *Chemical Abstracts* which comprise the CBAC data base are biologically oriented, computer searching of CBAC will identify only a limited number of articles relating to chemistry, regardless of the search strategy. Therefore, while computer-readable access to the chemical literature may be considered adequate for articles concerning biological activity, it should not be assumed that an exhaustive search by computer can be performed on the chemistry of a specific compound.

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 representative 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.