Development of Shell Chemical Company's Pesticide Information Retrieval System*

By JUDITH A. WIEMAN
Shell Chemical Company, Technical Information Services, New York, New York 10020
Received March 16, 1964

The Agricultural Chemicals Division is a wholly integrated division of Shell Chemical Company, which administers the manufacture and sale of pesticide chemicals and all attendant functions. In the Division office, one department has the responsibility for gathering the data necessary for submitting labels and petitions for tolerance. Another directs formulation development and toxiocological studies. Others are responsible for commercial development, process development, plant scheduling, possible patent interests, and packaging.

Within these groups, little data which are vital to a research facility are of need. For example, there is no need for data relating biological activity with functional groups and there is little interest in laboratory synthesis *per se*. The information which should be retrievable can be tabulated as:

1. Primary screening data—when and against what pests
2. Toxicological data—acute and long term
3. Performance data—including possible resistance and crop tolerance
4. Residue data—including analytical methods for determination, possible metabolites, and persistance
5. Formulation data—including ingredients, compatability, and synergism
6. Manufacturing data—including methods of synthesis, process design, and plant equipment
7. Packaging data—including suitable materials as well as storage stability
8. Marketing data—sizes of markets and information on competitive products

Sources.—The sources of this information are Shell Group Research and Technical Service Laboratory reports and performance reports from various State, Federal, and independent agencies which carried out field trials and toxicological studies.

Early Methods.—Two manual methods of information handling were tried and failed to retrieve information at a sufficiently efficient level. The first method was a simple subject heading system using host, pest, and compound headings. This method worked satisfactorily for short reports on performance data, but did not include the associated topics of toxicity, residues, tolerance, and resistance. And as reports became more voluminous, the method simply became too unwieldy. It had the additional drawback of using common names of pests which were not adequately cross-referenced. The second method tried was more comprehensive. McBee Keysort Cards were used. This could have been more successful than the subject headings except that it was inadequately implemented, causing lack of continuity and poor indexing practices in spite of the rules set forth. Furthermore, the categories were quite broad, and a great deal of actual document searching was still necessary, so this system was abandoned.

Development of Current System.—Technical Information Services is a company-wide, centralized group attached to the Head Office Engineering Department. It consists of three independent sections, interrelated by their common goal of information distribution, storage, and retrieval, operating under one supervisor. The three sections are:

1. Correspondence—staffed by a supervisor, three literature chemists, and five clerks
2. Library—staffed by a supervisor, three literature chemists, one literature biologist, and three clerks
3. Reports—staffed by a supervisor, five literature chemists, and three clerks

The Report Section, whose system is described herein, is the only section currently using mechanical methods.

Since the Report Section had already been indexing marketing, manufacturing, and packaging data for other Divisions of the company, we had certain category terms to cover these areas of the pesticide information. There were, however, no terms in the dictionary concerned with entomology, agronomy, or parasitology. These had to be selected.

In order to draw up a tentative word list, some 200 were randomly selected from the collection and read by the literature chemists. For clarification of some terminology, experts available in the Agricultural Chemicals Division were consulted. Terms were chosen on the basis of (1) frequency of appearance, (2) anticipated value in retrieval, and (3) acceptability as scientific terminology.[1] The list included names of major crops, geographical locations, terms to describe the scale of work being reported, the names of Shell's and competitive products, and the names of major pests. For indexing pests, the use of common group names such as moths, beetles, and so on, was impossible because of the imprecision of the nomenclature. The decision was made to use a strictly entomological classification by morphological families, usings as authority the "Bulletin of the Entomological Society of America," Vol. 6, No. 4. If, in the future, searching time becomes excessive owing to the use of families, further definition of pests by genus or genus species will be made.

* Presented before the Divisions of Chemical Literature and Agricultural and Food Chemistry, Pesticides Subdivision, Joint Symposium on "Problems of the Pesticide Literature and Some Solutions," 147th National Meeting of the American Chemical Society, Philadelphia, Pa., April 9, 1964.

(1) T. L. Gillium, *J. Chem. Doc.*, 4, 29 (1964).

Competitive product terms also caused difficulty. A conference on this topic was held at which the Division personnel stated that effort involved in solving the problem was not too great in comparison with ultimate results which could be achieved with a successful solution to the problem. The final solution arrived at was the inclusion of each product as it appeared in a document, cross-referenced common or trade names as they issued, and a triannual review of the use of these terms to determine whether they were valuable additions or designations for experimental compounds which had been dropped by the originator. If the latter were the case, the term would be deleted from the dictionary.

After modification of the tentative word list, trial indexing was run. Everything appeared satisfactory, so work was started to incorporate the word list into the existing dictionary and production indexing was begun.

IBM Card System.—The existing information retrieval system at this time was a modification of a procedure devised by Dr. Jack Sherman of the Texas Company.[2] The core of the method was a deck of 20,000 IBM cards, devised by Dr. W. F. Brown of Sun Oil Company. These cards were serially numbered in the first five columns and contained four random punches in the last forty columns. The four random punches became the code for the indexing category assigned to a particular card. When a report was indexed, it was assigned a serial number for identification purposes, and this number with certain title information were punched into the first forty columns of an IBM card. Using standard procedures, the four random punches of all categories by which the report was indexed were superimposed in the last forty columns of the card containing the serial number and title identification. This finished card was called the Literature Card. The Literature Cards were searched by sorter for the combination of punches corresponding to the categories sought. Since the code punches were randomly selected, the unrelated material was supposed not to exceed 5% when using as many as seventy category codes on one card. This was not so. The practical limit to avoid excessive false drops, when searching, was between thirty and thirty-five categories to describe one document. When indexing of the agricultural information was begun, it soon became apparent for indexing in depth that a maximum of thirty categories was intolerable. Also, the available number of categories was being used up at an alarmingly fast rate. These two items, coupled with the large volume of material to be indexed and subsequently searched, which would increase searching time by a large factor, gave rise to the consideration of the use of a computer.

After preliminary consideration of what was wanted in a computer program and after the economics were investigated and found to be favorable, the transformation of ideas into actuality was turned over to the Shell Information and Computer Services Department.

Computer Program.—The computer program, written in COBOL for use on an IBM 1410 computer, took approximately six months to write. The entire project consisted of two types of programs: two small programs to create the first Master Search Tape from data we had already stored on tape and three programs to update the master file once it was created, and the search program. The high density Master Search Tape is serial in form. Each document on the tape is a record of variable length. The first field is the assigned document number. The second is title identification information, and following these are the category numbers, in ascending order, by which the document was indexed.

The tape is updated by the following basic steps:

Input

1. A card with document number and title identification, followed by category cards for that document; one category per card with the associated document number. The category numbers are in random order at this point.

Processing

1. The input cards are sorted by category number.
2. The category numbers are passed against a tape, which we call the Category/Generic Tape, to pick up the categories which we have designated as generic in the category which was indexed.
3. The categories, now with associated generics, are sorted by document number, placed in ascending order, and written out, after the document number and title identification, onto the tape.

The update program allows corrections to be made to documents existing on the tape as well as the addition of new documents. Corrections may be: specific category additions to specific documents, specific category deletions from specific documents, blanket category deletions (*i.e.*, the deletion of a category from every document on which it appears), and last, the deletion of an entire document from the tape.

Current scheduling calls for the Master Search Tape to be updated once a week by the addition of 125 documents plus corrections. The factor limiting the number of documents is the manpower available for indexing. Searching time is provided once a day. As many as 36 searches may be performed, each having a maximum of ten categories sought, per pass of the tape. The searches may be only in two forms: A + B + C or A + B but not C. This allows a general search to be done concurrently, if desired, the results of which will not include the results of the specific search already sought and obtained. For example, the following search for information on the use of Product A to control house flies in barns can be set up:

Product A + houseflies + barns

If information of a more general nature would be helpful to the requestor, lacking the specific information, the search would be:

Product A + houseflies *but not* barns

Another example—in a search for data on the use of Product X to control aphids on lima beans, the search would be:

Product X + aphids + lima beans

and concurrently, for fringe or possibly helpful information search:

Product X + aphids + vegetables *but not* lima beans

(2) J. Sherman, "The Use of Four Hole Randomly Punched Cards for Abstracting Publications and Reports into IBM Cards," Texas Co., informal report, 1953.

The term vegetables is generic in the term lima beans.

The results of the first search in each example would answer a specific request, whereas the second would give possible helpful information which would not require examination of the information already found in the specific search.

Search time is scheduled for 2 P. M. each day. Searches received in the morning get "same day" answers. Those received after 1:45 P. M. are held until the next day unless there is an emergency. The search cards are punched by the literature chemist assigned to do the search. The card is verified, manually, by another literature chemist. The cards are delivered to the computer room, and when the run is completed the results are delivered by hand. The computer print-out of the search results has the following format: the header "The following documents satisfy the conditions of search ...." This is followed by a list of document numbers along with the title identification of the documents. The documents are screened by the literature chemist for pertinency, and all documents containing the requested information are given to the requestor for his evaluation.

A Working System for Retrieval of Chemical Structures, Adaptable to Pesticidal Screening Data*

By JOE R. WILLARD and EDWARD J. MALKIEWICH
Research and Development Department,
Niagara Chemical Division, FMC Corporation, Middleport, New York
Received April 2, 1964

The need for access to test data from past pesticidal screening, using as the basis for access the chemical structures of the compounds involved, led us several years ago to develop a structure retrieval system based upon a combination of empirical formula and functional group classification.

Because IBM data processing equipment was available to us, it was appropriate that any machine-sorting system devised be adaptable to this equipment. Further, every compound in our screening program is assigned a number at the time of its introduction into the program. This number, designated the NIA number, becomes an integral part of all the chemical and biological records and serves as an excellent "handle" for retrieval of information.

The Basic Worksheet.—The 80 columns of the IBM cards were divided into five areas:

Acquisition number (columns 1–6)
Elemental content (columns 7–13)
Structure classification (columns 14–68)
Company source (columns 69–70)
Biological data (columns 71–80)

Three pages make up the worksheet for coding a compound. The first two relate to chemical and source information and the third to biological information. For convenience, the 11-position of the column has been designated as X and the 12 or highest position as Y.

No 9-position appears in any column of our worksheet. When the cards are processed, the 9-position is punched in all columns where no other position is to be punched. This was originally intended as a quick means of checking the accuracy of punching. Other applications for this 9-position will be illustrated later.

Our worksheet is designed for direct coding in all areas except the company source area where reference is made to lists of companies corresponding to the numbers.

The acquisition number, or NIA number, is written on the code sheet (Fig. 1) and punched by direct reference. The numbers of carbon and hydrogen atoms in the molecule are also written directly on the code sheet. For elements other than carbon and hydrogen, the presence or absence is sufficient evidence to permit retrieval and is coded by circling the appropriate position of columns 11, 12, or 13.

NIA No. $\bar{1}\ \bar{2}\ \bar{3}\ \bar{4}\ \bar{5}\ \bar{6}$

	OTHER ELEMENTS		
	11	12	13
C	Y Ag	Y F	Y
	X Al	X Fe	X O
	0 As	0 Hg	0 P
$\bar{7}\ \bar{8}$	1 B	1 I	1 Pb
	2 Bi	2 K	2 S
	3 Br	3 Li	3 Sb
	4 Ca	4 Mn	4 Se
H	5 Cl	5 Mo	5 Si
	6 Co	6 N	6 Sn
	7 Cr	7 Na	7 Zn
$\bar{9}\ \bar{10}$	8 Cu	8 Ni	8 Other

Figure 1.

* Presented before the Divisions of Chemical Literature and Agricultural and Food Chemistry, Pesticides Subdivision, Joint Symposium on "Problems of the Pesticide Literature and Some Solutions." 147th National Meeting of the American Chemical Society, Philadelphia, Pa., April 9, 1964.