(5) Jaffe, H. H., and M. Orchin, "Symmetry in Chemistry," Wiley, New York, 1965.

(6) A similar approach was used by McDonnell, P. M., and R. F. Pasternack, "A Line-Formula Notation System for Chemical Compounds," J. CHEM. DOC. **5**, 56 (1965), but our subsequent treatment differs from theirs. See also Pasternack, R. F., and P. M. McDonnell, "Designation of Ligand Positions in Coordination Complexes," *Inorg. Chem.* **4**, 600 (1965).

(7) Cahn, R. S., C. K. Ingold, and V. Prelog, "Specification of Molecular Chirality," *Angew. Chem. Intern. Ed. Engl.* **5**, 385 (1966).

(8) Petrarca, A. E., J. E. Rush, and M. F. Brown, "A Nomenclature-Independent Method for Specification of the Stereochemistry of Coordination Compounds," Abstracts of Papers, INOR 147, 156th Meeting, ACS, Atlantic City, September 1968. (Also to be submitted for publication in J. CHEM. DOC.).

Storage and Retrieval of Agricultural Screening Data*

J. B. HAGLIND, H. J. ACKERMANN, R. E. MAIZELL,
T. M. MANNING, and B. S. SCHLESSINGER†
Technical Information Services, Olin Mathieson Chemical Corp.,
275 Winchester Ave., New Haven, Conn. 06504

Received September 23, 1968

A system is described for the storage and retrieval of agricultural screening data. The system embraces a formatted means of entering laboratory results in notebooks, keypunching the data onto cards, and processing the card file onto a disk for random access of data. A FORTRAN program is used to search the disk file for specific compounds and tests. Compound structure searching is achieved by using a permuted Wiswesser Line Notation file.

For several years prior to the development of the system to be described, biological test data determined by our Agricultural Research group had been recorded in conventional laboratory notebooks and keypunched onto cards for storage and retrieval. The cards were manipulated by standard unit record equipment and information printed from them by programs written for and operated on a 1401 IBM computer. This system was effective and satisfactory as a start. However, as the files of biotest data grew and secondary screening data rose in relative importance, a new approach was clearly needed (1) to store and retrieve not only current data, but also the large volume of previous data; (2) to reduce machine processing time and costs; and (3) to give greater flexibility and capability. We proceeded to try to meet these needs in closest possible cooperation with laboratory personnel.

ANALYSIS

Availability of an IBM 1130 and 1800 at the Olin Research Center in New Haven prompted development of a system suited for these computers and also adaptable, if necessary, to an IBM 360 scheduled for later installation.

It was decided that further processing of earlier results, already available on punched cards and in quarterly printouts, could await development of the system described here. The older files would then be integrated into the ongoing system as time and money permitted. Data could, of course, still be retrieved from older files by visual inspection of printouts or by sorting of the punched cards.

To simplify the system, it was agreed to use brief codes to indicate test results, rather than full descriptions. The input format was designed to contain the Olin Compound Registry Number, date of compound entry into the Registry, and 52 specific results for seven different test classifications (Figure 1).

DEVELOPMENT

The most important data, initially, were all primary screening information recorded in notebooks after a specified date. Since these data had been recorded in a relatively uniform manner in notebooks, selection of specific results for storage in the data retrieval system was simplified.

After deciding what information was to be stored, an input format was designed. A template was made so as to be superimposed on a notebook page to permit more rapid location of specific input for the system. Clerks manually transcribed data from notebooks to input forms for keypunch operators.

To eliminate the transcription step for current data, a unique notebook was designed to permit scientists to enter data in card column format in areas marked for rapid location by keypunch operators. Adequate space was given for additional comments or remarks (Figure 2).

* Presented before the Division of Chemical Literature, 156th Meeting, ACS, Atlantic City, N. J., Sept. 13, 1968.

† Present address: Department of Library Science, Southern Connecticut State College, 501 Crescent St., New Haven, Conn. 06515

Item	Columns
Card No.	1–2
Olin Compound Registry No.	3–8
Code No.	9–18
Date Received	19–22
Physical Form	23
Molecular Weight	24–27
Fungicide % Active	28–42
Herbicide % Active	43–60
Contact Insecticide % Active	61–68
Nematocide % Active	69–74
Aqueous Herbicide % Active	75–80

Figure 1. Agricultural primary screening input card format

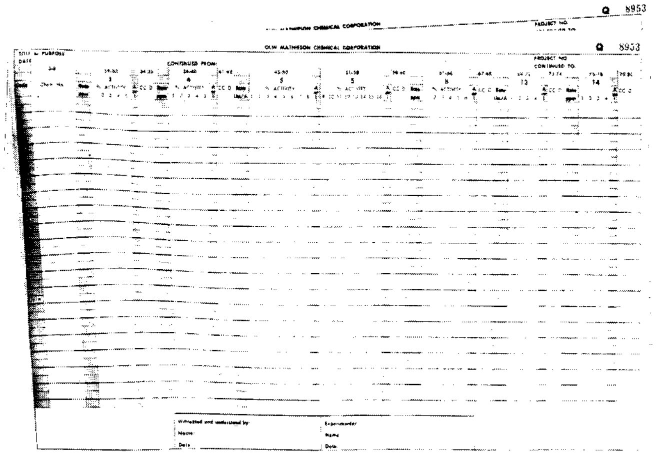

Figure 2. Research notebook used for agricultural primary screening data

The following sequence of steps was then followed: Carbon copies of completed notebook entries were removed from the newly designed notebooks and sent to keypunch operators on a current basis; data were keypunched and verified; and tabular listings were made via FORTRAN programs written initially for the 1130. New input data were also stored on a 2315 disk. FORTRAN programs were written to search the disk and to print responses to inquiries.

Upon installation of an IBM 1800, programs written for the IBM 1130 were easily adapted for the 1800, and data files were regenerated for this computer.

The data in the keypunched cards remain retrievable by use of mechanical sorters if the computer is not available, or if the search is specific and simple.

Retention of a copy of duplicate notebook pages permits rapid look-up of specific data directly from the input sheet, and also permits checks on accuracy of keypunched information.

CAPABILITIES

Use of full computer capabilities in the ongoing system permitted an appreciable savings in time over the previous manual sort techniques and computer print. Printouts could be produced on a regular schedule. Most important, however, was the more effective direction of the chemical synthesis effort which came about by the use of the revised notebooks and the printout of current screening results.

The FORTRAN programs written for the data retrieval system include:

1. Transfer of data from card file onto disk.
2. Processing the card file onto disk while checking the input for keypunch or technical errors—e.g., entry of criteria values which do not agree with those previously established in a set of standard guidelines.
3. Search for a specific compound number, or combinations of numbers up to 50 at one pass, for results with a stipulated cut-off date.
4. Search for a specific test result or combination of two test results.
5. Print of the entire file or portions thereof, depending on the operators' command or on requirements of the search question.

SEARCH ROUTINES

There are four types of search routines incorporated in the data retrieval system:

I. Single Parameter Type Inquiry Routine. This is used to search for one variable for one test on file. A typical question might be: "Are there any compounds which had a value of 6 or greater as a contact insecticide?" The computer begins by printing the following message:

```
AGRICULTURAL SCREENING DATA

XXXX RECORDS ON FILE

INQUIRY

DATA   INVENTOR   TEST NO. 9   CRITERIA 6
```

The XXXX represents the number of records being searched. The nine (9) indicates that the computer has been asked to look for the ninth test in the series. The six (6) is the criterion for the search—i.e., the computer will search for and accept every test with a level of performance equal to or greater than six. The computer stores every test meeting the requirements and then prints the message:

```
XXX HITS RETRIEVED
```

The selected print routine is written into the inquiry. Three options are available:

a. The computer prints out all of the data retrieved.
b. The computer prints out only the Compound Numbers for tests retrieved.
c. The computer does not print any data, but goes on to the next inquiry.

Option (b) is useful if a large number of tests are expected to be retrieved, and cuts down considerably on print time. Such a list permits use of microfilmed data sheets (which give compound number, molecular formula, and structure and property data) to screen compounds quickly. Option (c) was used with the IBM 1130 if the number of hits was larger than expected or if the individual chose not to pursue the question. The computer searches the complete file, printing each set of 50 hits. If the printed message on the number of tests retrieved indicates that less than 50 hits have been made, the file search is complete.

Zeroes are used as values in the file to denote that the compound failed the test. Blanks are left in the test record to indicate the compound was not screened by that test. The program permits bypassing blank data while searching.

II. Inquiry Routine Using "AND" Logic. In this routine, the computer accepts any tests that meet criteria set by the inquiry. A typical question could be: "Are there any compounds on file which have a value of 6 or greater as a Soil Fungicide and a value of 4 or greater as a Nematocide?" All specified parameters must be present in the record for a hit to be printed. Up to 50 test parameters may be set in conjunction with each other in a single inquiry. As in (I), blank data can be deleted if requested. The basic pattern of search and print options is the same as in (I).

III. Inquiry Routine Using "OR" Logic. In this routine, the computer accepts only those records that meet inquiry criteria which specify a number (up to 50) of disjunctive parameters—e.g., 1 or 2 or 3 or. . .*n*. A typical question might be: "Are there any compounds with values of 7 or greater as Aquatic Herbicides or Contact Insecticides or Nematocides?" Again, blank data can be deleted if desired, and the basic pattern of search and print options is the same as in (I).

IV. Inquiry Routine for Compound Number. This routine searches for Compound Numbers, to answer questions dealing with the existence of a specific compound or compounds in the file. The computer searches against the file for up to 50 numbers at one pass. It compares the file against Compound Number and prints out all data for each Number retrieved until all Numbers have been retrieved or the entire file has been read.

NEXT STEPS

When hits indicate promising compounds, additional data about these compounds may be obtained through other parts of the Olin Research Information System. This information may be retrieved through the Olin Compound Registry Number, since this Number is also used in the indexing of all internal technical reports.[1,2]

A permuted index of Wiswesser Line Notations (WLN) of Olin compounds permits rapid assembly of a list of all compounds related to those which have shown promise in the screening tests. As described by Granito *et al.*,[3] other WLN files may be merged with the internal file of Wiswesser Line Notations.

OTHER ACTIVITIES

The results from advanced stages of screening (secondary screening and field testing), can be readily incorporated into the system. Related information from research on analytical procedures, residue procedures, toxicology studies, and market data provide a system searchable for almost any combination of useful data in the development of new agricultural compounds.

Finally, we should mention that involvement of laboratory personnel in retrieval system development not only helped them to re-evaluate their work, but also to become familiar with the fundamentals of the retrieval system. In this way, they became more efficient users and contributors of useful suggestions for improvement.

ACKNOWLEDGMENT

Contributions to the development of this system by F. A. Dowell (formerly, Olin, New Haven), D. W. Ryker, and D. A. Shermer are gratefully acknowledged.

LITERATURE CITED

(1) Ackermann, H. J., J. B. Haglind, H. G. Lindwall, and R. E. Maizell, J. CHEM DOC. 8, 14–19 (1968).

(2) Schlessinger, B. S., and R. E. Maizell, "A New Approach to Indexing Technical Reports in an Industrial Information Center" (paper in preparation).

(3) Granito, C. E., J. E. Schultz, G. W. Gibson, A. Gelberg, R. J. Williams, and E. A. Metcalf, J. CHEM. DOC. **5**, 229–33 (1965).

Heuristic Retrieval: Variable Search Strategies for Identification

EUGENE S. SCHWARTZ
IIT Research Institute, 10 West 35th St., Chicago, Ill. 60616

Received September 12, 1968

Information retrieval is an empirically derived technique for identifying, locating, and retrieving specified information in a data file. The data file is a set of records each of which contains a description of an item in the form of numerical values or alphabetic descriptors. A set of values or descriptors constitute an information vector. A complex item can be described by a set of information vectors, each vector defining a different characteristic.

Given a data file with its sets of information vectors structured in coordinate and subordinate relationships, the problem of information retrieval is to isolate all items that match a specified description. The match is equivalent to satisfying the conditions of a Boolean expression.

Two types of strategy are generally employed in a search operation:

1. Sieve: selection by sorting on designated parameters in a described sequence.
   a. Positive sort: search for items in the file that have specific parameters.
   b. Negative sort: search for items in the file that omit specific parameters.
2. Interactive: open loop feedback between system and user.
   a. Parameter redefinition: results of a previous search are used to change search parameters.
   b. Relevance feedback: user "homes" in through question-and-answer procedure by adjusting the search requests to correspond to relevant items.