A Mechanized Storage and Retrieval System for Internal Documents*

WALTER G. MICHAEL
Philip Morris Research Center, Richmond, Va. 23261

Received May 8, 1973

A storage and retrieval system developed for pharmaceutical use has been adapted by Philip Morris Research Center for its internal documents. The system consists of a thesaurus of allowed terms and four files. Two important features include coordinate indexing using roles and links as well as the use of Wiswesser Line Notation for storing compounds. Hard copies of the thesaurus and files are computer produced, so that laboratory personnel can perform manual searches after a minimum amount of instruction. Questions too involved for manual searches can be approached by any of three mechanized routes: a document search, a Boolean search, or a mixed search.

The Philip Morris Research Center entered the mechanized information retrieval field in April 1962.[1] The initial system operated as a simple, card-sort process where fields were set up in a standard IBM card for different types of data, such as document number, date, and chemical name.

When this system became too cumbersome, a search was instituted for one that would hold input from laboratory notebooks, internal reports, and other Center-generated data.

A system was desired that would be usable by laboratory personnel after a short training session and would be easy to update. A survey of general requirements showed that a code-number system would be too difficult to teach to laboratory personnel and would become too cumbersome for the degree of specificity needed.[2] The desired system needed to accept virtually any information that could be described by words, symbols, or numerics.

THE SYSTEM—DESCRIPTION

It was agreed that the most expedient way to get such a system would be to acquire an information retrieval package which was already operating and had been debugged. Eventually, "The Combined File Search System," developed for the Food and Drug Administration by the Service Bureau Corporation in July 1968, was selected.

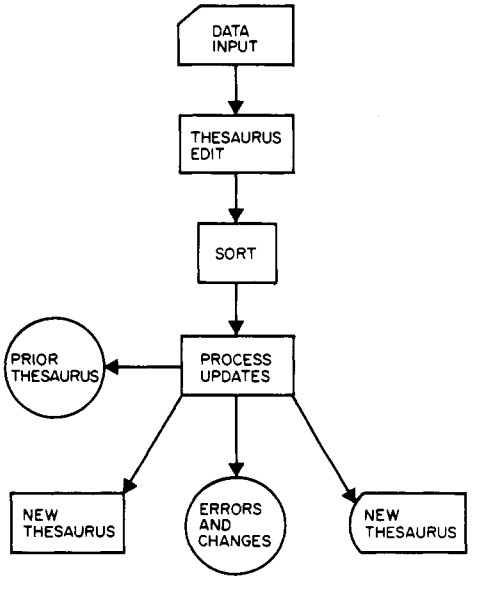
On request, the FDA provided a program tape, and a JCL deck was obtained from the Office of Air Programs at the Environmental Protection Agency Station in Research Triangle Park, N. C. Essentially, the system consists of a thesaurus of allowed terms and four files, the master file, inverted file, cross-reference file, and identification file.

The thesaurus is the backbone of the system. It is designed along the lines of the "Thesaurus of Engineering and Scientific Terms"[3] (Figure 1), using broader, narrower, see-also, used-for, and related terms. Each descriptor (term) is assigned at least one function (precise, common, or subdescriptor) and a subdictionary (author, chemical, Wiswesser, etc.). These subdictionaries permit the printing of separate subthesauri—e.g., a thesaurus of only razor-blade terms, chewing gum terms, or of tobacco terms, etc.

Scope notes are added to some terms to describe further the proper use of the term and also to give additional information such as Chemical Abstracts Service Registry Number, molecular formula, manufacturer, trade name, etc.

The thesaurus is also a validation device which checks new entries to ensure that only terms which have been approved are used to describe a document or search request. The first phase of this thesaurus function is the edit. The input data are checked to assure that all new terms have a function and that appropriate subdictionaries are assigned. The edit program then converts the data into a suitable processing form, prints the errors encountered and incorporates the edited data into the previous thesaurus.

Another program then prints changes, additions, deletions, and errors encountered in this merging phase and also posts the broader-narrower terms and the use-used



THE THESAURUS VALIDATES AND EDITS TERMS USED IN A DOCUMENT UPDATE OR A SEARCH AND PROVIDES A LISTING OF ACCEPTABLE TERMS.

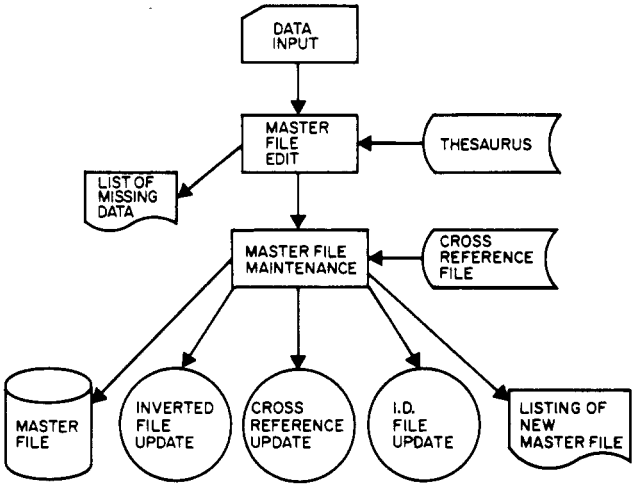
Figure 1. Thesaurus

*Presented in part before the Division of Chemical Literature, ACS, March 14, 1973, Columbus, Ohio.

for relationships. The thesaurus also has a subroutine for subordinate term extraction. This is a process that separates all subordinate terms in the thesaurus and sorts as to particular type, such as use terms, see-also terms, related terms, etc. These extracted terms are sorted into alphabetical order and validated against the thesaurus to ensure that all subordinate terms are also listed as main terms. Any that are not will be detected and listed for correction.

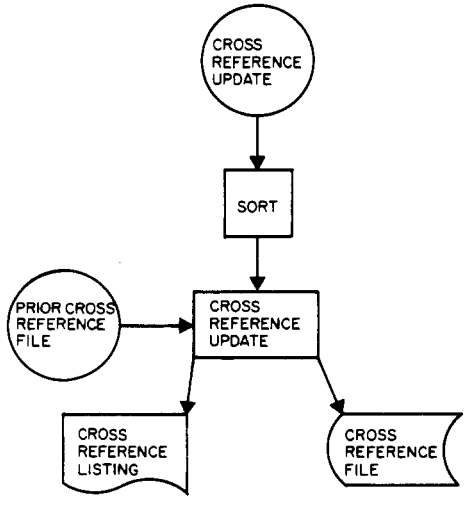
The next component of the information system is the master file (Figure 2), which contains documentary information, including keywords, titles, notes, and indexer-selected data. This master file is used to build the other three files.

The Research Center master file is divided into two sections. The first one contains all documents written from 1970 on. These documents have been keyworded by an indexer. The second section is comprised of all documents written before 1970. These documents have been keyworded by their authors, with no quality control standards applied. The additions and changes to the master file are



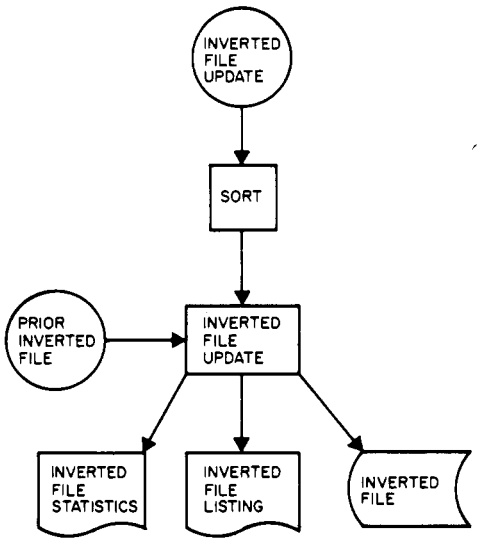
THE MASTER FILE CONTAINS ALL INFORMATION FROM WHICH THE OTHER THREE FILES ARE BUILT AND IS THE BASIS OF THE RETRIEVAL SYSTEMS.

Figure. 2. Master file



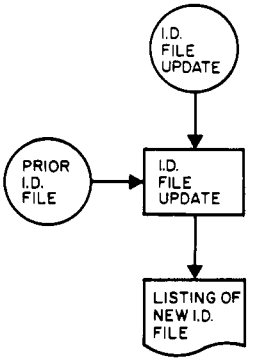
THE CROSS REFERENCE FILE RELATES INTERNAL, COMPUTER-ASSIGNED DOCUMENT NUMBERS TO NOTEBOOK AND REPORT NUMBERS.

Figure 3. Cross-reference file



THE INVERTED FILE IS AN INDEX TO THE MASTER FILE

Figure 4. Inverted file



THE IDENTIFICATION FILE IS A CONDENSED MASTER FILE CONTAINING DESCRIPTORS SPECIFICALLY TAGGED AS I.D. FILE ENTRIES (BROADER TERMS, AUTHORS).

Figure 5. Identification file

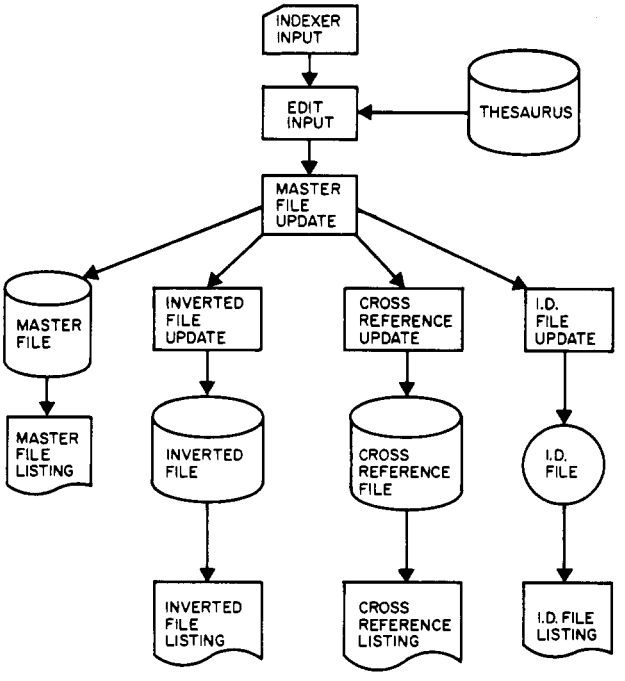


Figure 6. System flow chart

```
SEMICARBAZIDE,MONOHYDROCHLORIDE
  FUNCTION-- PD          DICTIONARIES--4

  BROADER TERM--
    ZVMZ &GH
  SEE ALSO--
    RT SEMICARBAZIDE
  SCOPE NOTES--
    CAS. REG. NO. 563417        CH5N3O.HCL

SEMICARBAZIDE,THIO
  FUNCTION-- PD          DICTIONARIES--4

  BROADER TERM
    ZYUS&MZ
  SEE ALSO--
    RT SEMICARBAZIDE
    RT THIOSEMICARBIZIDE HYDRACHLORIDE
  USED FOR--
    N-AMINOTHIOUREA
  SCOPE NOTES--
    CAS. REG. NO. 79196         CH5N3S

SENEOL
    FUNCTION-- PD        DICTIONARIES--4

  BROADER TERM
    L6UTJ CQ1OVR& DO2& BO1

SENEPOXYDE
  FUNCTION-- PD          DICTIONARIES--4

  BROADER TERM--
    T36 BO DUTJ FO2& GO2& A1OVR

SENINE
  FUNCTION-- TD          DICTIONARIES--4

SEPARATION
  FUNCTION-- PD          DICTIONARIES--2


  SEE ALSO--
    NT DIALYSIS                 NT EXTRACTION
    NT FILTRATION               NT FLOTATION
    NT REMOVAL

SEPARATORS
  FUNCTION-- PD          DICTIONARIES--2

  BROADER TERM--
    EQUIPMENT

SEPARATORS
  (CONT.)
  SEE ALSO--
    NT CYCLONES                 NT SIEVES

SERINE
  FUNCTION-- PD          DICTIONARIES--4

  BROADER TERMS--
    AMINO ACIDS                 QVYZ1Q
  SCOPE NOTES--
    NONESSENTIAL AMINO ACID
    CAS. REG. NO. 56451         C3H7NO3

SETTLING RATE
  FUNCTION-- PD          DICTIONARIES--2

SEX
  FUNCTION-- PD          DICTIONARIES--2

SHARP, G H
  FUNCTION--PD,ID        DICTIONARIES--3

SHAW, W S
  FUNCTION--PD,ID        DICTIONARIES--3

SHEET
  FUNCTION-- TD,SD       DICTIONARIES--2

SHEET WEIGHT
  FUNCTION-- PD          DICTIONARIES--2

SHIFFLETT, A W
  FUNCTION-- PD,ID       DICTIONARIES--3

SHIPPING
  FUNCTION-- PD          DICTIONARIES--1

  USE--
    FREIGHT TRANSPORTATION

SHORTS
  FUNCTION-- PD          DICTIONARIES--2

  BROADER TERM--
    TOBACCOS
```

Figure 7. Thesaurus

EXTERNAL / INTERNAL	EXTERNAL / INTERNAL	EXTERNAL / INTERNAL	EXTERNAL / INTERNAL	EXTERNAL / INTERNAL
NB5248-73	NB5277-23	NB5371-01	NB5388-08	N4557A-01
00001185	00001028	00000790	00001382	00001271
NB5248-76	NB5277-30	NB5371-03	NB5388-14	N4557A-25
00001183	00000813	00000789	00001381	00001076
NB5248-77	NB5277-54	NB5371-11	NB5388-71	N4557A-96
00001184	00000791	00000787	00001368	00001096
NB5248-86	NB5277-75	NB5371-25	NB5388-96	71-0046
00001300	00001024	00001236	00001364	00000323
NB5277-02	NB5277-78	NB5371-81	NEW NO	71-0048
00001026	00001243	00001011	00001195	00000989

Figure 8. Cross-reference listing

```
                    INVERTED FILE LISTING                    PAGE 41

MALIC ACID
    42, 381, 829, 914, 979, 1161, 1162

MALIC ACID,DIPOTASSIUM SALT
    379, 383

MALIC ACID,DISODIUM SALT
    379

MALONIC ACID
    42, 979, 1303, 1384

MANGANESE
    372

MARKET SURVEY
    322

MARLBORO
    307, 323, 356, 389, 403, 761, 764, 772, 776, 778, 796, 799, 800, 803,
    809, 845, 846, 848, 849, 887, 890, 891, 892, 897, 898, 899, 997, 1003,
    1012, 1020, 1025, 1037, 1039, 1057, 1088, 1089, 1176, 1204, 1206, 1269,
    1277, 1280, 1281, 1284, 1341, 2978

MARLBORO LIGHTS
    405

MARTENSITE
    406, 1038

MARTIN, P G
    63, 316, 326, 328, 999, 1099, 1222

MASS SPECTROMETERS
    333, 335, 917, 1303, 1305, 1309, 1311, 1312, 1314, 1316, 1325

MASS TRANSFER
    53, 344, 870
```

Figure 9. Inverted-file listing

validated for correct operations, spelling, contents, and sequences. Valid inputs receive any broader term or use term posting and are then accepted into the master file. Invalid inputs are printed along with explanatory error notes for correction. Any error causes the entire document to be rejected. This precludes the GIGO syndrome.

When a document is accepted into the master file, it is assigned an internal document number by the computer as shown in Figure 3. This number increases by one every time a document is accepted. This gives a numeric order to accepted documents, allowing for a random order of input. To allow the requestor to search the files by original document number—i.e., a number which refers to the researcher's original notebook or report, a cross-reference file is maintained which correlates the computer-assigned number with the original document number. The master file edit program uses the cross-reference file to reject duplicates which may try to enter the system.

Figure 4 describes the third file—an inverted file. This file is an alphabetical listing of descriptors along with a tabulation of all internal document numbers in which the index terms are used. The inverted file is the main tool for performing manual searches.

Figure 5 describes the identification file, which consists of broader terms, authors, and dates.

Two handy tools used in the Research Center system are coordinate indexing and Wiswesser Line Notation. The coordinate indexing system used is the Engineers Joint Council–Battelle approach of assigning roles and links to index terms. The Wiswesser Line Notations are

treated in the same manner as index terms and are assigned roles and links just as a chemical name is assigned its roles and links. Because the WLN's are also listed as broader terms, the computer will assign the appropriate WLN when the indexer encodes a chemical name. Permuting by rings has proved satisfactory and is the only degree of permuting being considered at this time.

Figure 6 summarizes the essential structure of the system.

USING THE SYSTEM

The next group of figures clarifies some of the definitions by illustrating printouts created by the system.

Figure 7 is page 142 of the thesaurus.

Figure 8 is a page from the cross-reference listing. This run picked up a duplication of notebook number 4557, and therefore N4557A was assigned to the second notebook issued as 4557 to eliminate this error. Although the system was not desinged primarily to detect errors of this type, in doing so, it gives a means for checking the filing system.

Figure 9 is page 41 of the inverted file. This lists the index terms used in the system by alphabetical order followed by the appropriate internal document numbers. If one wanted to do a manual search on malonic acid and mass spectrometers, for example, he would simply compare the internal document numbers and find that 1303 is common to both; therefore, document 1303 is a hit.

Figure 10 is a copy of the inverted-file statistics. This lists the descriptors in alphabetical order, followed by the number of documents in which this descriptor has been used, a statement about the last time the descriptor was updated, and/or deleted, and a statement of the function which the term fulfills in the thesaurus.

The Research Center System can perform four basic types of searches: document search, Boolean search, mixed search, or manual search.

The first three of these search types involve use of the computer. A manual search will normally be limited to use of the inverted file. For machine searches, there is a general list of options built into the search program. These are listed in Table I.

INVERTED FILE STATISTICS 09/15/72 PAGE 13

DESCRIPTOR	COUNT	LAST ADDITION	LAST DELETION	FUNCTION
DODD, C G	4	09/18/72		PRECISE
DODECANE	1	05/07/72		PRECISE
DODECATRICONTANE	1	09/18/72		TEMPORARY
DOTRIACONTANE	4	09/18/72		PRECISE
DOTRICONTANE-C14	8	09/18/72		PRECISE
DOVER, I C	13	09/18/72		PRECISE
DRAGENDORFF REAGENT	2	05/07/72		PRECISE
DRAWINGS	18	09/18/72		PRECISE
DRIVERS (AUTO)	1	01/23/72		PRECISE
DRYERS	1	05/06/72		PRECISE
DRYING	15	09/18/72		PRECISE
DUNN, W L, JR	1	08/28/72		PRECISE
DYTOL	22	09/18/72		PRECISE
EDMONDS, M D	43	09/18/72		PRECISE
EDWARDS, W B	16	09/18/72		PRECISE
EFFICIENCY	2	05/06/72		PRECISE
EICHORN, P A	1	01/23/72		PRECISE
ELECTRIC POTENTIAL	5	05/06/72		PRECISE
ELECTRODES	2	05/06/72		PRECISE
ELECTROLYSIS	4	08/28/72		PRECISE
ELECTROLYTIC CELLS	4	08/28/72		PRECISE
ELECTRON MICROSCOPES	5	09/18/72		PRECISE
ELECTRON PHOTOMICROGRAPHS	4	09/18/72		PRECISE
ELECTRON SPIN RESONANCE	2	09/18/72		PRECISE

Figure 10.

Table I. Search Options

(1) Specify maximum number of responses to be printed
(2) List only total number of "hits"
(3) Print, or not, index terms and text segments of document hits
(4) Truncation
(5) Specify numeric values
  a) equal
  b) not equal
  c) greater than or equal
  d) greater than
  e) less than or equal
  f) less than

LITERATURE CITED

(1) Murrill, D. P., *J. Chem. Doc.* **2**, 225-8 (1962).
(2) Murrill, D. P., *Lab. Management* **5**(6), 18-21 (1967).
(3) "EJC Thesaurus of Engineering and Scientific Terms," Engineers Joint Council, New York, N. Y., March 1969.

A Conversational Mass Spectral Search System. IV. The Evolution of a System for the Retrieval of Mass Spectral Information

STEPHEN R. HELLER,* RICHARD J. FELDMANN, HENRY M. FALES, and G. W. A. MILNE
National Institutes of Health, Public Health Service, Bethesda, Md. 20014

Received April 10, 1973

A prototype of an interactive, conversational mass spectral search system, developed at the National Institutes of Health, has been tested since September 1971 and is now being used by more than 200 scientists in the U. S. and Canada. The response has led to management of the system being given to the Mass Spectrometry Data Centre, Aldermaston, England for use by the international mass spectrometry community.

Over the past few years the Division of Computer Research and Technology (DCRT) at the National Institutes of Health has been developing prototype components of a Chemical Information System (CIS) for use by chemists and biomedical research personnel at NIH. Included in this work has been research on computer generation of Wisesser Line Notation (WLN),[1] sequential[2] and nested tree[3] structure searching, manipulation of three-dimensional structural data,[4] NMR data retrieval[5] and analysis,[6] and mass spectral data retrieval.[7-10] The last of these proj-

* To whom correspondence should be addressed at: Heuristics Laboratory, Division of Computer Research and Technology, Building 12A, Room 3001, National Institutes of Health, Bethesda, Md. 20014