NBS

Technical Note 285

FILE ORGANIZATION FOR A LARGE CHEMICAL INFORMATION SYSTEM

RUTH ANDERSON, ETHEL MARDEN, AND BEATRICE MARRON

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

THE NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards is a principal focal point in the Federal Government for assuring maximum application of the physical and engineering sciences to the advancement of technology in industry and commerce. Its responsibilities include development and maintenance of the national standards of measurement, and the provisions of means for making measurements consistent with those standards; determination of physical constants and properties of materials; development of methods for testing materials, mechanisms, and structures, and making such tests as may be necessary, particularly for government agencies; cooperation in the establishment of standard practices for incorporation in codes and specifications; advisory service to government agencies on scientific and technical problems; invention and development of devices to serve special needs of the Government; assistance to industry, business, and consumers in the development and acceptance of commercial standards and simplified trade practice recommendations; administration of programs in cooperation with United States business groups and standards organizations for the development of international standards of practice; and maintenance of a clearinghouse for the collection and dissemination of scientific, technical, and engineering information. The scope of the Bureau's activities is suggested in the following listing of its three Institutes and their organizational units.

Institute for Basic Standards. Applied Mathematics. Electricity. Metrology. Mechanics. Heat. Atomic Physics. Physical Chemistry. Laboratory Astrophysics.* Radiation Physics. Radio Standards Laboratory:* Radio Standards Physics; Radio Standards Engineering. Office of Standard Reference Data.

Institute for Materials Research. Analytical Chemistry. Polymers. Metallurgy. Inorganic Materials. Reactor Radiations. Cryogenics.* Materials Evaluation Laboratory. Office of Standard Reference Materials.

Institute for Applied Technology. Building Research. Information Technology. Performance Test Development. Electronic Instrumentation. Textile and Apparel Technology Center. Technical Analysis. Office of Weights and Measures. Office of Engineering Standards. Office of Invention and Innovation. Office of Technical Resources. Clearinghouse for Federal Scientific and Technical Information.**

---

*Located at Boulder, Colorado, 80301.
**Located at 5285 Port Royal Road, Springfield, Virginia, 22171.

NATIONAL BUREAU OF STANDARDS

Technical Note 285

ISSUED APRIL 18, 1966

FILE ORGANIZATION FOR A LARGE CHEMICAL INFORMATION SYSTEM

Ruth Anderson, Ethel Marden, and Beatrice Marron

Institute for Applied Technology
National Bureau of Standards
Washington, D.C.

For sale by the Superintendent of Documents, Government Printing Office
Washington, D.C., 20402 - Price 25 cents

TABLE OF CONTENTS

FILE ORGANIZATION FOR A LARGE CHEMICAL INFORMATION SYSTEM

Ruth Anderson
Ethel Marden
Beatrice Marron

ABSTRACT

The report describes a file structure which combines list-processing concepts (for handling variable length information records) with standard serial record arrangements (for identification information). The file organization was designed for a large chemical information system and includes both well-structured and unstructured (amorphous) information. An investigation was made of representative data inputs from the Department of the Army. The data to be put into the file, the nature of the file structure, and the necessary programs for manipulation of file information have been considered as interdependent parts of a total system. Computer programs have been initiated to test the validity of the proposed approach.

Key Words:
file organization, chemical information, chemical structures, linear notations, list processing, threaded lists, heirarchical files, master files, satellite files, structured files, level codes, pointers.

1. INTRODUCTION

This report describes a new approach to the structuring of a large file containing diverse information. The Army Research Office sponsored the research here described in support of its requirements for a large integrated chemical information system.

The Department of the Army has over five hundred installations which handle technical information; approximately one third of them handle information related in some way to chemistry. Many of the laboratories handle chemical structures, but there is a diversity in the other kinds of information associated with the structures; that varies according to the mission and requirements of the particular laboratory. The project here described has been concerned with an attempt to devise a file structure which will permit the storage and retrieval of all categories of information contained in any of the Army's laboratories (not all of which exist in any one laboratory) in association with the structure of a chemical compound.

Initial investigations were concerned with exploration of file organization research undertaken by others in an effort to determine whether techniques developed for different situations could be applied to the organization and manipulation of chemical information. Samples of CIDS "hard-core" data were obtained from various laboratories and a procedure for organizing files of this information was devised. This procedure combines the features of a fixed-length sequential master record file with those of a variable-length non-sequential data file strung together in a list-processing fashion. Present work includes programs for file creation, updating and expansion. Future plans include the creation of sub-files and the querying of the files.

2. OBJECTIVES

The project has been concerned with two considerations; long-range objectives and short-term goals.

The long-range objectives are:

1. To define some of the characteristics of large files, and to develop the structure for a large file of heterogeneous scientific and technical information including chemical structure representations, with particular attention to the necessity for:

    a. Manipulating information which is in some cases formatted and in others completely amorphous.

    b. Making provision for inclusion of certain kinds of information when it exists, and for later filling of gaps when the information is not available at the time of file initiation.

c. Creating a multi-level file of information so that provision may be made for the inclusion of generic information, as well as the addition of various levels of specificity as required by the user, either simply because the additional levels of specificity exist and might be useful at a later date or because there is a user requirement for that degree of specificity.

d. Examining the inter-structuring of files that are part of the larger files but which may be geographically separated.

2. To provide list-processing capability in the file structure in order to:

a. Maintain flexibility in the file.

b. Provide for an efficient means of updating.

c. Permit additions to files, both in classes existing already within the system and in the entry of new classes of information.

d. Free the system from the constraints of fixed-length and formatted files.

e. Permit aggregations of data from files that are geographically separated.

3. To investigate the techniques of file manipulation in order to provide systems of sub-files, special files, desirable redundancy in exchange for multiple access to information, and the necessary keying or cross-referencing facility required for such a system of multi-level, multi-subject files. This work must take into account also the planning and development of the several kinds of computer programs which are required for file maintenance. (This report presupposes that all such file manipulation will be carried out by computers.)

4. To determine the kind of organization of information which will most readily permit questioning of the file by different groups of questioners who have varied (and varying) requirements for the kinds of information contained in the file, and who have, furthermore, requirements for differing degrees of specificity in the information they are seeking.

The short-term goals are concerned with attempting to satisfy the following immediate requirements:

1. To store specific categories of well-defined data in a computer.

2. To devise procedures for questioning the file and retrieving information from it.

3. To employ computer search programs on a limited basis to test the validity of the approaches.

3. BACKGROUND STUDIES

Theoretical studies and actual implementations of files developed by others were investigated with a view toward incorporating those approaches deemed applicable to large chemical information systems. All of the file organizations studied were designed for structuring information for automatic storage and retrieval, but were responsive to different requirements. The studies investigated included variations of list-processing techniques, hospital and clinical record keeping, and chemical structure files containing associated corollary information. Appendix A lists the general background material for the development work here described; our work was a departure from any of these particular systems because of the specific nature of the Army's requirements.

The classical concept of a master file containing all known information for each entry in full detail was considered, but rejected because of its requirement for large amounts of space and because of its restraint on expansion in as yet unknown directions. Similarly, an investigation was made of the idea of a "Control File" or "File Relater and Locator System" which would consist of a separate file of identification numbers of entries and the names of the satellite files wherein information on these entries is located, together with the keys for addressing them. After consideration of the types of data to be included in a large chemical information file (see Section 4), it was decided as a first approach to devise a system incorporating the features of the two plans discussed above, i.e., a linear file consisting of basic fixed-field information about the entry, plus pointers to satellite information files.

4. TEST DATA

The U. S. Army (see Appendix A-1) lists the following eight types of information to be included for each compound in their chemical information and data system:

1. Registration or identification number
2. Chemical structure, probably a listing of atoms and bonds
3. Molecular formula
4. Bibliographic citations
5. Nomenclature, including chemical names, linear notation, trade names, etc.
6. Location of data files information
7. Kinds of data and information available at each location
8. Security classification and releasability

One other item not listed is the source of the compound or material that is the subject of the file entry.

The Army's file is designed to accomodate the entry of three million compounds. It should be noted that the above list includes both structured and unstructured information and that the list includes referrals to satellite files.

The Army was asked to supply sample data for the experimental attempts to structure the files, and in answer six organizations supplied sample forms which had been executed in accordance with a set of instructions supplied by the Army (see Appendix B). Some of the test data suppliers were Army laboratories and others were organizations who had contracts with the Army. The major portion of the information was recorded on the CIDS Registry File - Hardcore Input Forms (SMUEA Form 13, 26 AUG 64). (See Appendix C.)

As might have been expected, there was considerable variation not only in the content of the information supplied on the sample forms, but in the completeness of the individual entries. Certain of the organizations employed forms of their own design rather than those shown in Appendix C. There were differences in interpretation among the various

organizations as to what was desired by the Army for certain categories of information. All of the forms were incomplete in some respect, with some having little more than a structure diagram and the name of the reporting agency. Based on the use of the trial forms, and as noted in the Conclusions (see Section 7), additional effort should be expended on the design of forms and the development of instructions for completing them in order to promote greater consistency.

Personnel at Frankford Arsenal compiled a list of 1008 different categories of information in an attempt to identify all of the information requirements of the Department of the Army. This list will remain open-ended for the subsequent addition of new categories as need arises. Any one installation would not contain all of the elements of information which appear in the complete list. An analysis was made of the information received from the six sources listed in Appendix B, and it was found that the information supplied on the forms contained only a small number of the categories from the list compiled by Frankford Arsenal. No attempt was made to relate them directly to the Frankford list, but it was observed that the elements of data encountered on the forms tended to be formed into different classes rather than to comprise sub-sets of the Frankford listing. It can be assumed that the data reported responded to the requirements of the individual reporting agency.

For the sake of the experiment, it was desirable to treat all of the data from the six reporting agencies in a uniform way. Therefore, the following composite list of all the data reported was formed; no one set of forms contained all of the categories, and some contained only a few.

1. Reporting agency
2. Security classification
3. Release restrictions
4. Local control number
5. Date
6. Registry number
7. Molecular formula
8. Nomenclature
9. Structure

10. Bibliographic references

11. Types of data

12. Key words

Notations (such as Wiswesser notation, Hayward notation, IUPAC cipher, etc.) were included under nomenclature. These should form a separate category since nomenclature also includes trivial names, trade names, chemical names, common names, and others. (See Conclusions, Section 7). With respect to certain other categories of information, some forms merely listed the presence of information in that particular category, without supplying the actual information itself.

5. PROPOSED FILE ORGANIZATION

A tentative file organization to handle the categories of information contained on these forms is now under development. It will serve as an experimental model for the purpose of ascertaining from the proposed users its expected utility for their requirements.

The overall file system will consist of two parts:

1. A master-file of fixed length information,

2. Information-files of variable length information.

Further, the master file will contain keys to the location, size and type of pertinent entries in the information files.

Each record in the master-file will be assigned a unique identification (ID) number and each item in the information file will be tied to its master-file record by the ID number and will be identified by a "level code" to indicate its hierarchy in a two-dimensional chained-file arrangement. The following list of such categories was used as an experimental model, but is by no means an all-inclusive one. (Note assigned "level code" at right.)

Table 1

LIST OF FILE CONTENTS BY CATEGORY

Category	Level Code
1. ID number	
2. Reporting laboratory	
3. Local control	
4. Security	
5. Date	
6. Molecular formula	010000
7. Notations	020000
a. Hayward	021000
b. Wiswesser	022000
8. Nomenclature	030000
9. Types of data	040000
a. Physical properties	041000
b. Chemical properties	042000
c. Physiological effects	043000
1. Respiratory	043100
2. Cardiac	043200
3. Neuromuscular	043300
d. Toxicity	044000
1. Intravenous	044100
2. Intramuscular	044200
a. Rabbits	044210
b. Rats	044220
3. Oral	044300
10. Provision for supplementing categories in the above listing	

Table 2 is a machine word schematic for the fixed-length master-file record.

Table 2

SCHEMATIC FOR FIXED-LENGTH MASTER-FILE RECORD

Word no.		
1.	ID number	identification data
2.	Reporting laboratory	
3.	Local control number	
4.	Security data	
5.	Date	
6.	Pointer to molecular formula record	pointers to information files
7.	Pointer to notations record	
8.	Pointer to nomenclature record	
9.	Pointer to types of data record	
10.	Unassigned	

The fixed-length master-file record will occupy 10 machine words. Words 1 through 5 consist of identification information. Words 6 through 10 are "pointer" words. They serve as indicators of the presence or absence in the information file of their respective categories of information with a key to location and length of record when presence is indicated.

A typical pointer-word consists of three fields: one field contains the level code for a particular category of information; the second field contains the address where that particular block of information can be found; the third field reveals the automatically-generated number of machine words of data contained in that information record. Certain addresses in this field of a pointer word imply special conditions about the referred to data. This will be illustrated with specific information shortly.

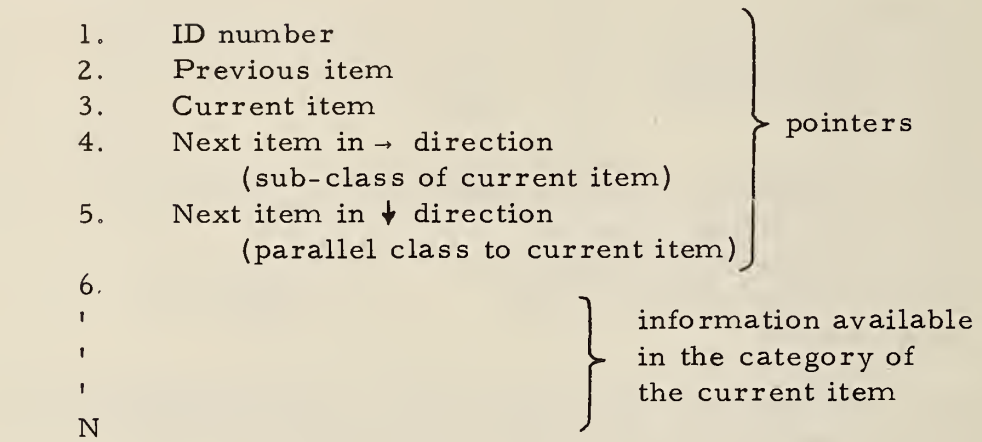
Pointer Word

level code	address	number of words

Table 3 is a machine word schematic for a variable-length information-file record.

Table 3

SCHEMATIC FOR VARIABLE-LENGTH
INFORMATION-FILE RECORD



The variable-length information-file record contains five fixed words at the beginning of each record: the first word is the ID number (which is shown as word 1 in the master-file record in Table 2); the next four words are pointers; word #2 points backward to the previous item in the hierarchy of levels, word #3 identifies the current item and words #4 and #5 point forward allowing a branching in two directions. The actual locations of the blocks of data are irrelevant, as long as the pointers link them correctly. The basic filing algorithm is the following: each information record filed is placed in a block whose address is in an index register called OPEN. This address becomes the address of the current item being filed and the index register is adjusted so it contains the address of the next OPEN block.

Let us consider a master-file record typical of those submitted, where information is given under each category listed in the data outline (Table 1) except for nomenclature, level code 030000.

Table 4

TYPICAL MASTER-FILE RECORD

Word no.	Contents				Explanation
1.	A0000007				ID number
2.	001500				Reporting laboratory code
3.	0006000				Local control number
4.	000000				Security information
5.	120764				Date
6.	010000	MOLE	07	POINTERS	Molecular formula
7.	020000	HAYW	03		Notations
8.	000000	0000	00		Nomenclature
9.	040000	DATA	09		Types of data
10.	000000	0000	00		Provision for additional categories

Words 6, 7, and 9 indicate that information pertaining to their respective level-codes can be found at the named machine locations. For example, word #6 indicates that seven words of information for level-code 010000 (molecular formula) can be found at location MOLE. Word #8 is zero, indication that there is no information in the file for level-code 030000 (nomenclature). Word #10 is blank indicating that no additional categories of information have been added to the master-file record.

Note how the above master-file arrangement leads itself to cursory scanning of "pointers" for the purpose of creating sub-files for each major category. Further illustrations will demonstrate that this facility for creating sub-files can be carried out quickly and efficiently at any given level in the file.

Let us now look at location MOLE for a typical Information File item.

Table 5

TYPICAL INFORMATION FILE RECORD - EXAMPLE 1

Location		Contents		Explanation
MOLE		A0000007		ID number
MOLE + 1		MASTER	POINTERS	Pointer to previous item
MOLE + 2	010000	MOLE 07	POINTERS	Current item
MOLE + 3	000000	0000 00	POINTERS	Next item →
MOLE + 4	000000	0000 00	POINTERS	Next item ↓
MOLE + 5 thru MOLE + 11		The molecular formula information		

Note from Table 1 that molecular formula is a one entry level. Therefore words MOLE+3 and MOLE+4 are pointer words of ZERO indicating the end of the chain. MASTER in MOLE+1 is a special symbol to point back to the master file. Since the master file will be sorted and rearranged, it was decided not to use the original address in pointers back to the master file. Note that words at MOLE and MOLE+2 are identical with words 1 and 6 respectively in Table 4.

Now let us consider the representation of an information record in a more complex network such as intramuscular toxicity. (See Table 1, category 9. d. 2, Level Code 044200.)

Table 6

TYPICAL INFORMATION FILE RECORD - EXAMPLE 2

Location		Contents		Explanation
INMUS		A0000007		ID number
INMUS+1	044100	INVEN 09	POINTERS	Previous Item
INMUS+2	044200	INMUS 00	POINTERS	Current Item
INMUS+3	044210	RABB 05	POINTERS	Next Item →
INMUS+4	044300	ORAL 11	POINTERS	Next Item ↓

The zeros in the third field of INMUS+2 indicate that the current item is a stepping stone to items further on in the hierarchy. Here there is no general information on intramuscular toxicity, but there is

information on intramuscular toxicity in rabbits located at RABB. Further there is information on oral toxicity located at ORAL.

Note that the level code of the next item → and the level code of the next item ↓ (here words INMUS+3 and INMUS+4) both can be converted to the level code of the current item when the following rule is applied: Starting from the right, decrease the first non-zero digit in the level code by one. This will prove useful in threading backwards in the hierarchy!

6. CURRENT STATUS

Information has been punched on 8-channel teletype tapes in ASCII (American Standard Code for Information Interchange) from the data mentioned earlier in this report. A program has been written for creating the master-file records from these data tapes. In addition several programs already operational are available from other projects for manipulating files. These programs, written for the NBS Chemical Structure Manipulation Projects, are as follows:

1. A system for a chemical structure or substructure search based on the Hayward notation.

2. A sort routine for sorting blocks of information on a specific key in the block.

3. A program for the derivation of molecular formulae and other structure-related screening information from the Hayward notation.

Programs are being written for transforming both Hayward and Wiswesser notations into connection tables of the same format. Generic structures of the Markush type and partially indeterminate structures will be interfiled with such connection tables. A program is also being written for a chemical structure search from the connection tables of both specific and Markush structures.

An executive routine is planned for the system which will include modules for file maintenance, creation of subfiles, and the exercise of certain control functions.

7. CONCLUSIONS

Based upon the initial research efforts reported here, the following activities appear to provide fruitful areas for continuation of the research in the organization of large chemical files.

1. Continue efforts to find ways of representing unstructured information. The development of a systematic classification scheme may be necessary.

2. Make provision for indicating in the automated information file that additional information is available elsewhere, and indicate where it is located; e.g., microfilm files, hard-copy folders at the reporting laboratory, or other.

3. Design forms for reporting information, along with a comprehensive set of instructions for proper recording of the information on the forms. Uniform reporting on standard forms will tend to promote greater consistency in the input data. This is an especially important consideration if a central file is to be built which will serve several agencies.

4. Amend the CIDS list of hard-core items to include notation as a separate item rather than as a part of nomenclature.

5. Continue the development of programs for entering information into the files and for searching the files.

6. Develop a package of software containing service routines and executive routines which will exercise control over file maintenance and selection of appropriate search routines for the file.

Appendix A

BIBLIOGRAPHY

1. U. S. Army Chemical Information and Data System, Status Report, February 1964 (AD432000)

2. J. C. Shaw, A. Newell, H. A. Simon, and T. O. Ellis, "A Command Structure for Complex Information Processing" in "Proceedings of the Western Joint Computer Conference, Los Angeles, Calif., May 6-8, 1958" (American Institute of Electrical Engineers, New York, N. Y., 1959), p. 119 ff.

3. Feasibility of Establishing an Integrated Agency Wide Scientific Information System(s). Report to FDA, DHEW April 1964 (C65851) Arthur D. Little, Inc.

4. Hospital Computer Project, Memorandum #5 (#1075) December 31, 1963, Bolt Beranek and Newman, Inc.

5. The Multi-List System. Technical Report No. 1, November 30, 1961, The Moore School of Electrical Engineering (AD270572. 3)

Appendix B

TEST DATA SUPPLIERS

	Name	No. of Forms
1.	U. S. Army Biological Laboratories Fort Detrick (Disinfectant files)	49
2.	U. S. Army Biological Laboratories Fort Detrick (Crop files)	222
3.	Edgewood Arsenal	50
4.	Aberdeen Proving Grounds	29
5.	Chemical Abstracts Service	100
6.	University of Pennsylvania	105

CIDS REGISTRY FILE - HARDCORE INPUT

1. REPORTED BY				2. COMPOUND SEC. CLASSIF.
				3. GROUP
5a. LOCAL CTL. NO.	5b. CTL. CHANGE	5c. DATE MO. DY. YR. 12 17 64	6. REGISTRY NO.	4. RELEASE RESTRICTIONS

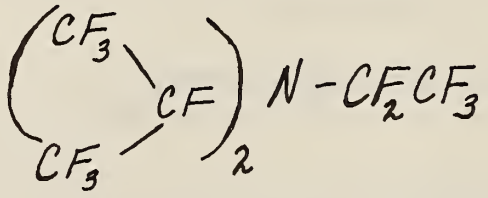
7. MOLECULAR FORMULA

$C_8F_{19}N$

8. NOMENCLATURE

Trisdylamine, tridecafluoro -1,1 - bis(trifluoromethyl)
Ethylamine, N,N-diisopropyl -, perfluoro-
Diisopropylamine, tetradecafluoro -N- pentafluoroethyl-.

9. STRUCTURAL FORMULA



LOCAL CTL. NO.

CIDS REGISTRY FILE - HARDCORE INPUT CONTINUATION SHEET

10. KINDS OF DATA

Boiling point
Index of refraction
preparation

11. BIOGRAPHICAL CITATIONS

46:06015 p Kouck & Simons Brit. 666, 733

47:08772 b Kouck et. al.
p U.S. 2.616.927 11/4/52

U.S. DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20230

OFFICIAL BUSINESS

POSTAGE AND FEES PAID
U.S. DEPARTMENT OF COMME