Processing Data from a Large Drug Development Program*

E. H. ECKERMANN,** J. F. WATERS, R. O. PICK,*** AND J. A. SCHAFER
Division of Medicinal Chemistry,
Walter Reed Army Institute of Research,
Washington, D. C. 20012

Received October 25 and December 30, 1971

Almost three million records covering approximately 175,000 chemical compounds and involving multiple disciplines have been recorded in the Army Antimalarial Drug Program. Maximum compartmentalization of files together with a complete method of interfacing files and a unique number system has made it possible to manage this system.

The Army Antimalarial Drug Development Program was designed to test large numbers of compounds. To decrease the probability of missing an active compound, several types of screens were established. These multiple test systems not only allowed the program to take advantage of proven procedures developed by the effort generated during World War II, but permitted the rapid exploitation of new ideas and methods. This also made it possible to drop several less selective test systems without adverse effect on the over-all program.

We anticipated that up to 100,000 compounds would be examined with as many as 10 records per compound required to record the screening data. In addition, inventory and chemical data had to be maintained for each compound. In fact, chemical data proved to be more massive than biological data in terms of storage. To date, approximately 175,000 compounds have been tested. Approximately two million biology screening records, 850,000 inventory records, and 600,000 chemical records have been generated. As many as 10,000 records have been added to the master files in a single week. The chemicals received for testing have been submitted from over 1000 sources. Since the data serve each submitter as a guide for further synthesis, they must be forwarded as soon as possible. Many of these submitters have participated in this program under an agreement in which they are assured that all information regarding their submissions will be released only to them or to previously cleared individuals having need of the information for further evaluation. This restriction made it imperative to be able to associate the data rapidly, not only for a given compound, but for a specific lot of a compound, with a specific submitter. Such a massive data file obviously could only be managed by computer and the computer facilities at WRAIR were completely inadequate to meet these requirements. Subsequently, the problem of designing a system to handle these large and diverse files was complicated by the necessity to plan for the use of computers operated by other organizations.

*Contribution No. 1032 to the Army Research Program on Malaria.

**Present address: Office of the Assistant for Veterinary Affairs, Forrestal Building, Washington, D. C. 20314. To whom correspondence should be addressed.

***Present address: Department of Medical Research & Development, William Beaumont General Hospital, El Paso, Texas 79920.

SYSTEM ORGANIZATION

The system which was designed to manage the data has undergone changes as the program developed. However, the main functions and the basic concepts remain. The three major divisions of the data system (biological, inventory, and chemical) were designed so that they could be maintained and used individually, but could also be used together in various combinations. It is completely feasible to maintain these three divisions on three different computers, and, in fact, the inventory was maintained first on an IBM 360 and later on a CDC 3300 while the other two were maintained on an IBM 7090 and an IBM 7094.

The interface of the divisions of the system is based on identification of data by bottle number. Each inventory and biological record that enters the system contains the bottle number. The bottle number corresponds to a lot and sub-lot number. This number is assigned to a compound when it is received, allowing the compound to be processed immediately. Later the compound is also given a compound identification number which corresponds to a chemical structure and applies to all lots of a given structure. When the program first started, the compound was assigned the compound identification number and the lot was identified by a suffix. However, this led to a considerable delay in processing the compound while it was checked against prior submissions for duplication. If an error were made, corrections had to be made on all files. Under the present system, the bottle number is used for identification and the association of a bottle number to a compound identification number is made only on the inventory file. In this way, errors need to be corrected only on the inventory file. There is no need to correct other files which contain only the bottle number.

BIOLOGY FILES

Computers can be a very helpful tool, but the investigator often finds himself a servant of the computer, having to organize his work and record his data to conform to existing systems. We decided that in this program every effort would be made to avoid unnecessary inconvenience to the investigator. The investigator determines how he wishes

to record biological data. A recording sheet, from which data can be keypunched directly, is then designed to accomodate each laboratory. The only requirement is that the compound being tested be identified by a bottle number on each record. Some investigators have found it to their advantage to utilize ADP for their own use. In these cases, we utilized their software to the maximum extent possible. As a result of this policy, inputs to the biological files are received in various formats and stages of processing.

The great variation among the various biological files made it essential that they be maintained separately for updating and editing. Originally it had been intended to merge the biological files together as early as possible in the processing, but the rapid growth of the individual files soon made it apparent that a merged file would be extremely cumbersome. Thus, the biological files are compartmentalized, although formats were standardized as much as possible in order to utilize standard merges and sorts.

In addition to making an otherwise cumbersome biological file easily manageable, compartmentalization has provided considerable savings and flexibility in the system. Selected files can be used individually for searching or may be merged with other selected divisions of the system. Processing can thus be broken into shorter jobs which better utilize machine time and allow more than one machine to be used. Compartmentalization proved to be such an advantage that it is now used to the maximum possible extent in the entire system.

When biological data are received, they are reformatted and maintained on tape as 84 character records blocked 10. When reformatted, the first eight characters contain the bottle number, the next four contain the test date, and characters 79 and 80 contain the test system code. The rest of the record contains the biological data and varies appreciably from one test system to another. These files are all maintained in bottle number sequence. After merging with inventory files, biological file records are expanded to 102 characters and maintained 102 characters blocked 10. The first 25 characters contain common data as follows:

1–8 Compound Identification Number

9–15 Bottle Number

16–19 Source Number which is an Identification Number of the Individual or Organization submitting the compound

20–25 Sort Key

The bottle number is present in both biological and inventory premerge records. In merging the biological and inventory files, the compound identification number and source number are located in the inventory record and duplicated in the corresponding fields in the expanded biological record. The compound identification number, bottle number, and source number fields contain the information most commonly used for searches and represents, in our opinion, the maximum amount of data that should be inserted in all records in order to facilitate searches. Since any portion of the biological and/or inventory file may be examined on any search, random access does not appear to offer any great advantage over serial files.

INVENTORY FILES

There are two sub-files in the inventory division of the system. The first contains compound data such as chemical name, chemical and physical properties, and condition of the sample. Administrative data such as the bottle number, compound source, shelf location, date received, amount on hand, and cross reference data such as submitters' identification number and compound identification number. This file is maintained on disks and indexed by bottle number. It currently has about 350,000 records.

The second file contains data on individual shipments. Included in each record is the bottle number, information on the screening center, the test performed, the date of shipment, and the date of receipt. This file is also maintained on disks and has about 500,000 records.

Output of 375 characters blocked 10 may be produced for merging with one or more biological files. The identifying numbers of these outputs correspond to those of the biological files. Upon merging with biological files, an output of inventory data is produced for further processing with chemical files. These files contain 120 character records.

CHEMICAL FILES

The chemical files have proved to be the most difficult to handle.[1, 2] While programming has been divided into manageable segments, it has not been possible to divide the files into units that might logically be used independently of each other. Sub-files have been maintained for use in special situations, but they cannot be used together as a substitute for the master file. By maintaining chemical files independent of other files, it would be possible to substitute into this system with a minimum of effort any other chemistry file in which the compounds are identified with our compound identification number.

The chemical file contains basically four areas: (1) A BCD area with molecular formula and identification number (currently the bottle number). (2) A bit-coded portion consisting of a molecular formula and screen. The screen in this context is a bit which can be turned on or off and reports a specific pre-asked question about the structural characteristics of the compound. (3) The third area is also bit coded and reports a topological map of the compound, atom by atom and bond by bond. (4) Output codes to print a structural formula on a special drum printer.

Structure queries are input via the chemical typewriter. Full structure and substructure matching is done on first an "at least" basis. That is, the file compound must contain at least those screens, atoms, and bonds contained in the query. Once this criterion is met, atom-by-atom comparison is done to determine either identity or inclusion.

There are several characteristics of the chemical file which present problems during the interface with files from biological and/or inventory files.

(1) The file must be maintained in molecular formula sequence to facilitate efficient searching and updating of the file.

(2) Since no duplicates are maintained on the file, not all lot numbers will be present on the file. (This situation is discussed in the Interface and Use of Files.)

(3) Owing to variations in the size of chemical structures, variable length records are necessary.

INTERFACING AND USE OF FILES

When a compound is received, it is immediately assigned a bottle number. This number is also provided to the submitter and investigators testing the compound. After processing through the chemical data files to check for duplication, a compound identification number is assigned. Since this number identifies the chemical structure, a compound may be assigned a new number, or if the compound is a duplicate of a compound already on file, it is assigned the same number as the compound which it duplicates. A separate, unique suffix identifies each lot and salts are identified informally for drug purposes.

Therefore, a compound has only one compound identification number, but may have several bottle numbers. The structure for this compound will appear on the chemical file only under the first number accessioned. (This number will later be referred to as the structure accession number.) This number is now a bottle number, although earlier other accession numbers were used. Catalog numbers and the number the submitter gave to the compound are but two examples of other accession numbers used. Biological data may be interfaced with chemistry and printed with the chemical structure for distribution. More frequently, however, it is printed after merging with inventory in order to reduce processing time and to reduce bulk. This merge provides a printout with the lot number, chemical name, source number (which identifies the submitter), submitter and laboratory identification numbers, and biological data. For internal distribution to the staff, the compound identification number is also included. Data are usually printed in two sequences: compound identification number sequence for in-house use and in source number sequence, which may be divided easily and mailed to submitters. In source sequence, the compound identification number is not printed, and the submitter receives information only on the lot of the compound that he submitted.

Interfacing biological data with chemical data is a complex operation. Post-merged biology and inventory records are expanded to 174 characters. The interface allows for expansion of data to a print format, the addition of a field for line control and the addition of an identification number which is common to chemical, biological, and inventory files. A common number is necessary since a structure is identified on the chemical file only under the bottle number which was originally assigned (the structure accession number). Therefore, if a second bottle of a compound is received, that bottle number must be related to the number on the chemistry file, the structure accession number. A file of synonomous numbers is maintained in bottle number sequence on a separate file. This is the means by which structures of duplicate compound identification numbers but differing only in suffix may be associated with the proper biological data. This file is matched with inventory and/or biology files and the structure accession number duplicated in a new field in the inventory and/or biology file. All files (chemical, inventory, biological) are then sorted into structure accession number sequence. At this point, any combination of one to three biology and/or inventory files may be merged. This number was determined by the configuration of the computer. Any of the three of these files may contain two or more merged individual biology and/or inventory files or outputs of searches of a file or files. To facilitate searching of files for activity or structure function relationships any one of these three files or the chemical file may be designated as a master to extract information for corresponding compounds from the other files. The output from this merge may then be sorted into compound number, bottle number, or source number sequence before printing. Since the data have already been expanded to print format, the print program is a simple one. It does require a special printer, however, in order to print the special characters required for the chemical structure. Until the present time, it was practical to edit and format the individual small files before the final merge and to use a simple generalized print program to generate output for the printer. For small searches and for small files this is still a practical method. However, with the rapid increase in the size of the files, a point has been reached where the sacrifice of this method for one large, sophisticated, post-merge program appears necessary in order to carry shorter records through the sorts and merge. This change is being undertaken.

The system has undergone changes as the antimalarial drug program became more complex. Three changes mentioned earlier have been vital to the continued use of this program:

(1) The immediate assignment of a bottle number to each chemical submission with later identification by compound identification number

(2) Maximum compartmentalization of files to allow separate maintenance and use

(3) Development of methods to allow for interfacing of the various parts of the system

This system was designed for the antimalarial drug program, but has been easily adapted to the radiation drug and schistosomiasis drug programs. We believe that it would be equally adaptable to other programs of this nature.

ACKNOWLEDGMENT

We acknowledge the technical assistance of Gen Jue and Daniel Boehle and the editorial assistance of Mildred Garrison and Peggy Casteel.

LITERATURE CITED

(1) Feldman, A., Holland, D. B., and Jacobus, D. P., "The Automatic Encoding of Chemical Structures," *J. Chem. Doc.* **3,** 187 (1963).

(2) Jacobus, D. P., Davidson, D. E., Feldman, A. P., and Schafer, J. A., "Experience with the Mechanized Chemical and Biological Information Retrieval System," *J. Chem. Doc.* **10,** 135 (1970).