Edited by Joe J. Hanan

Bulletin 417
March, 1985

Published by the Colorado Greenhouse Growers' Assoc., Inc. in cooperation with Colorado State University

research bulletin

USERS' INTRODUCTION TO DATA BASE

Charles W. Basham[1]

Every organization, enterprise, or project needs an information system. If the enterprise is limited in scope the information system may be quite informal, relying heavily on the memory of an individual and a few books or pieces of paper hardly deserving of the term "file". As the enterprise gets larger or more complex or as more than a single individual is involved, the need for a more comprehensive approach and greater formality is quickly felt. A grower can probably recall from memory the plant characteristics and general productive capability of each crop, but memory probably is unreliable as to specific cropping history, fertilizer treatments, and other such details, particularly if several greenhouses and crops are involved. The usual solution to this problem is to get some filing cabinets and set up a filing system.

[1]*Associate Professor, Department of Horticulture.*

Paper Files

Paper files are in fact simple, widely used, and well understood tools for managing data. A paper file system also requires a relatively small initial capital investment. The investment of time and effort in maintaining and using such a system may, however, be considerable and it is a continuing operational cost to both owner and user.

The involvement required by owner and user (the same individual is both owner and sole user in the simplest case) is depicted in Figure 1. The activities involved in file maintenance and file search are not trivial in any case and can become overwhelming with increased diversity in files, users, and owners as diagrammed in Figure 2.

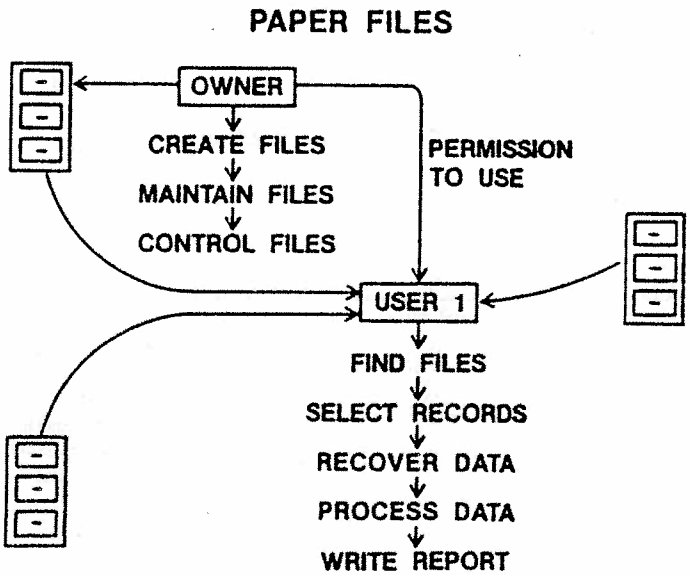

Figure 1

This bulletin is published in cooperation with Colorado State University Experiment Station and Cooperative Extension Service. The information given here is supplied with the understanding that no product discrimination is intended and that no endorsement of a product is implied.

PAPER FILE ENVIRONMENT

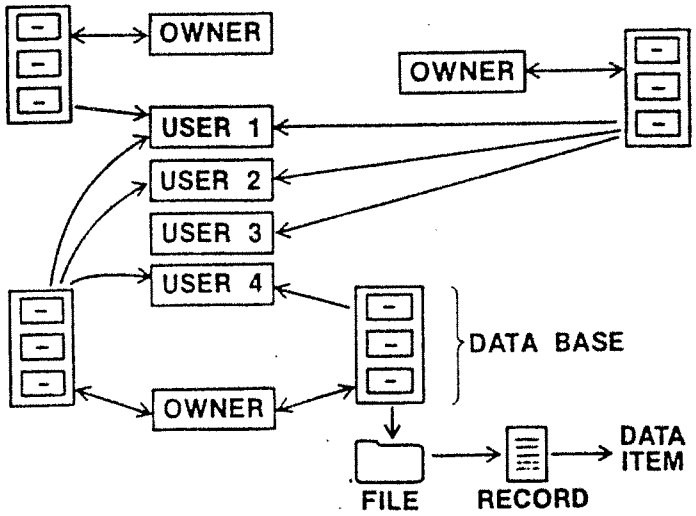

Figure 2

A Note on Terminology

Figure 2 represents a data base as a cabinet of files. The simplest definition of a data base is from Martin, 1981: "a collection of data which are shared and used for multiple purposes". There is no mention here of the media on which the data are stored (paper or electronic). From the point of view of a user, the data base may consist of a few files in a single storage unit or many files, with diverse ownership and location, to which that user has access permission. A person interested in land use might conceive of a shelf of soil survey reprints as a data base or part of a larger data base extended to include aerial surveys showing land use patterns.

The concepts of file-record-data item are dependent, i.e., related, A data item is the smallest unit of data that has meaning, the smallest unit of named data. In a plant pedigree, the name or registration number of a variety is a data item. A "record" is composed of a group of related data items, e.g. a pedigree can be viewed as a record consisting of plant identifiers (data items — names and registration numbers); the relation is established by the structure showing parent-progeny relations over several generations. The concept of a file is a collection of related records, e.g. the pedigrees of all the cultivars in a specie.

Some of the problems with paper file systems are:

1. There is a continuing requirement for rather large inputs of time and effort on the part of the owner/user. Maintenance, updating, and retrieval take considerable time. If maintenance and updating are slighted the system will quickly become less and less useful.

2. Retrieval of information on any but the primary key (a key is a data item that identifies a data grouping) is difficult. The history of fertilizer applications to a crop might be retrieved from fertilizer invoices containing data on fertilizer analysis, amount, application date, and field, as well as other data items. If these invoices are filed under "fertilizer" or "crop history" the retrieval is relatively simple; if they are filed under supplier names or "paid bills" the retrieval may be very difficult and time consuming. Developing a report relating fertilizer history and yield history would introduce further retrieval difficulties.

3. The user must go to the physical location of the file. Unless the files are centralized this may mean going to several locations to gather data required for a single report.

4. Simultaneous searches by multiple users may be difficult or impossible. It is really not practical for two or more persons to search through the same file cabinet at the same time. Even if the searching is not truly simultaneous a previous searcher may have temporarily removed some files for copying.

5. There may, probably will, be a high degree of redundancy and inconsistency in the data. Data items that are nominally the same may appear in different records and different files within the system (redundancy); the values recorded for these data items may differ from record to record (inconsistency). The amount of fertilizer to be applied to a crop, for example, may appear in a fertilizer recommendation, the amount ordered for that crop may appear on a purchase order (and it may differ from the recommendation), and the amount delivered may appear on an invoice (and it may differ from the amount ordered). There is the possibility then that three fertilizer use reports could be generated, each of them showing different amounts.

6. Physical space and facilities may become limiting as the files grow. Without some method for removing and archiving dated material the cost of physically accommodating the files may become quite large.

Electronic Files

For the past several years there has been a steady migration of information systems from paper files to electronic files with computer control and processing. The concept of file and record is basically the same in electronic and paper files although the physical entities are quite different. In electronic files the data items are coded in a machine readable form, most commonly on magnetic disks which serve as peripheral mass-memory storage devices for the computer (Figure 3).

A Note on Computer Systems

A computer system can be thought of as three basic components interacting to accomplish a task:

1. **Hardware** — composed of all the mechanical and electronic devices such as terminals, central processor, and peripheral memory (disk drives) necessary.

2. **Software** — composed of coded instructions that control the activities of all the hardware components and direct those activities toward accomplishment of a given task.

3. **User** — the person who defines the task and selects the appropriate combination of hardware and software to achieve results.

The user, while listed last here, is obviously central and critical to the system. Without an intelligent, knowledgeable, responsible user, the system is aimless.

The data items are entered into the computer memory in file form, i.e., by records structured into files in a format that preserves and defines the logical relations in the data. This structured entry of the data is done through a software interface of some kind, most often a program that is a part of the computer's operating system. This same program can be used to retrieve a listing of the data entered into the file, but to generate any other kind of report requires an application program supplied by the user. For example, to generate a report listing all the progeny of a given cultivar from a file of pedigrees, an application program to access the proper file and search for progeny of that cultivar and to write a listing of the progeny found would be necessary. Since the logical relations between data items in records and records in files are imbedded in the organization of the data in memory these relations must be specified in the application program.

Many, perhaps most, of the problems in an electronic file environment can be related to the dependency between data and application programs sketched above. Figure 4 indicates the complexity of an electronic file environment though there may be a great many more users, application programs, files and owners in a real system than diagrammed here, and the dependencies and inter-relations may be much more complex. File management systems are available as software packages and are imbedded in some packages such as SPSS (Statistical Package for the Social Sciences). It seems fair to say that, while these file management routines provide more or less help in dealing with the problems, they do not solve the problems.

Some of the problems with electronic files, in summary are:

1. As with paper files, there tends to be a high level of redundancy and inconsistency of data in the files. Since data files are often created to support a specific application, redundancy is almost enforced. Inconsistency springs from the same sources as for paper files and from the fact that it is unlikely that all files are updated on the same schedule.

2. The system is inflexible; since the application programs must carry information on file structure and extraction of data from files is dependent upon application programs, any new information requirement is likely to require a new application program and it may require a new data file.

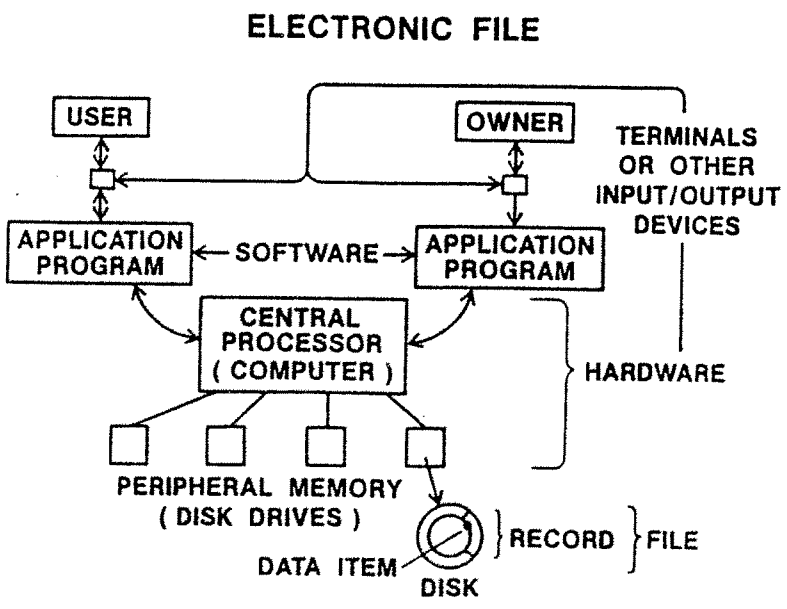

Figure 3

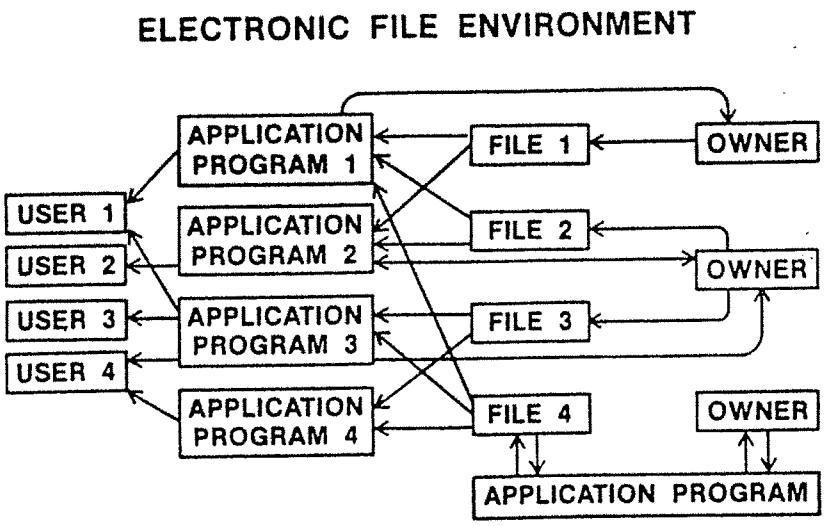

Figure 4

3. It is expensive to change; any change in data, application program, or hardware configuration is likely to require a cascade of changes through the system.

4. All the above problems are based, to at least some extent, on the mutual dependence between data and application programs.

Data Base Management Systems

A *data base management system* or DBMS is the collection of software programs required for using a data base that resides in a computer and consists of a collection of related data available on a controlled basis to multiple users and their application programs without requiring user or application program to know how the data are organized in storage. Conceptually, the DBMS can be segmented as in Figure 5. The definition processor structures the storage scheme and positions data in storage so it may be retrieved. The query processor allows the user to access logical records and files from the data without an application program. It also provides special facilities for access to the data by application programs, a description of the data structures no longer needs to be included in each program.

A Note on Files and Records

We have just introduced the idea of logical files and records; previously we have talked of physical entities when speaking of files and records. In paper and, to a large extent, electronic file environments the logical structure is reflected in the physical arrangement of files and records; e.g. pedigrees (records) put together in a file folder (cultivar file). In a data base environment the logical structure of the data is held in the software and the physical location of the data is meaningless to the user. The user or application program can request files and records of the desired logical structure through the query processor and they will be con-

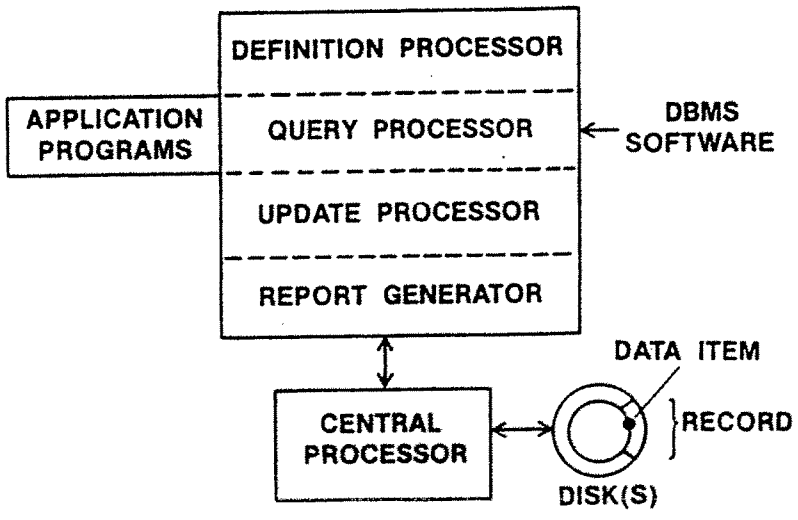

Figure 5

structed from data items, whose physical location and relationships in the data store is known only to the DBMS, and returned to the user as logical files and records. In short, a logical file is that file perceived by the user and has no relation to the physical location of the data.

The update processor allows for additions, changes, and deletions of data items. The report generator is especially useful in that it typically allows the user to do arithmetic and logical transformations and selections and to get a report in suitable format containing the results. For example, the user might request, and get, a list of all the progeny of a given cultivar whose current age is greater than 3 years and less than 6 years, which had a yield in excess of a given minimum (assuming of course that the necessary data are present in the data base). Such a query requires no application program.

Some of the characteristics of a data base environment are diagrammed in Figure 6. A new actor is introduced on the scene here; the data base administrator. This person (or team) is responsible for controlling data entry, including updates and other changes, and user access. The role of the owner has disappeared from the scene.

A Note on Ownership

The dictionary says to own is to have or hold as property. In the case of paper and electronic files, an owner has been identified with the files. The owner may hold the file as property, or as a member or employee of an organization, may have ownership in the sense of responsibility for and authority over the file while property rights reside in the organization, not the individual. In a data base environment, ownership may continue either in the sense of property or of responsibility and authority. The rights of ownership, however, are now exercised through the data base administrator. The owner may indicate which of the users may have access to particular information and it is the responsibility of the data base administrator to enforce those stipulations through the security provisions provided in the DBMS.

The rights, responsibilities, and authority of ownership still exist, but they are behind the scenes. They are exercised through the office of the data base administrator.

Some of the desirable characteristics of a data base management system environment are:

1. The physical file has disappeared; logical files are created for the user by the DBMS upon request. There is an appearance of ownership to the user for all data to which that user has access permission.

2. Redundancy is controlled through design and administration of the data base. Many different users and application programs may require the same data item within differing logical files, but the data item generally needs to be stored only once. This data item is inserted in the proper format into each of the requested logical files that are then delivered to the user.

3. Inconsistency in data is reduced or eliminated. If each data item is stored only once, then that single value will appear on every report calling for that item generated from the data base.

4. The data structure and the application programs are independent; either can be changed without affecting the other. In a properly designed and executed data base the basic problem of electronic file systems is cured. The data base is subject centered while the file environment is application centered. The data base should be able to accommodate the data requirements of any properly coded application program, always subject to the presence of the requisite data in the data base. In a file environment it is common for each new application program to require creation of a new file and it is almost certain that changes in files will require changes in application programs.

5. The DBMS can respond much more easily to requests for information in an unanticipated form; depending upon the kind of request made and the specific design of the

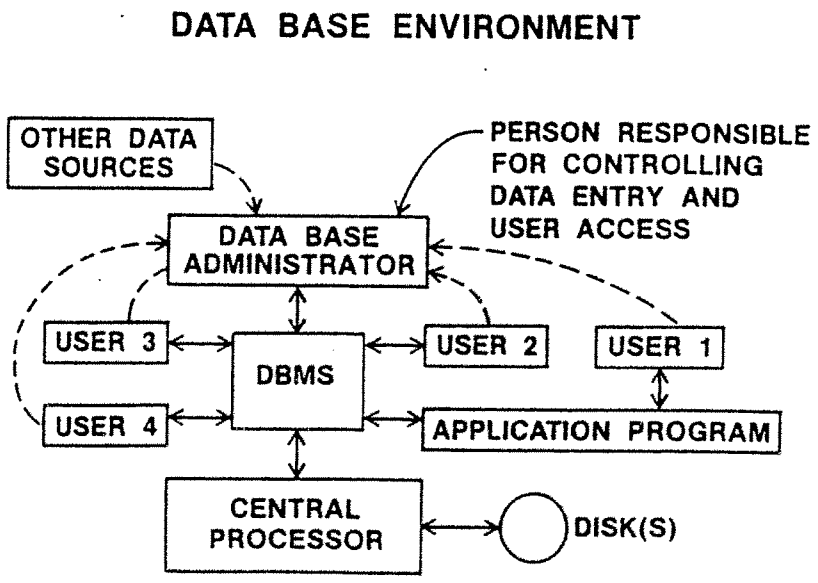

Figure 6

data base, some results will require greater processing time than others. In a file environment, response to unanticipated requests are often simply not feasible within any practical constraints of time and cost.

6. The security and integrity of the data are controlled through design features of the DBMS and through actions of the data base administrator. Security is controlled through access permission; the data base administrator may give a user permission to only certain segments of the data base and then only to read data, not to modify it. There may be several levels of permission enforced through passwords and other mechanisms invoked by the administrator. Data integrity is protected by strictly controlling modification privileges. DBMS software also commonly contains routines for error checking as data are entered into the system.

7. The system can supply data and generate reports without an intermediary application program. A wide range of user requests, even those requiring complex data transformations, sorting, classifying, and selection, can be serviced directly by the DBMS with no need for an application program. Where an application program is required it is typically easier to write and maintain because of the special accommodations built into the DBMS.

A Final Note of Caution

Data base technology is an exciting area of computer applications with tremendous potential impact. The cautions listed below apply to any source of data, not just to reports from a DBMS. However, because many of us are likely to be functioning in a DBMS environment in the future they are worth emphasizing.

1. The fact that you retrieve data from a data base does not assure validity, reliability, accuracy, or relevance. The appearance of data on a computer report confers no special status. The user must look to the original source of the data to make quality judgments and always be wary of mistakes that somehow get in between the original source and the report in hand.

2. A data base system may encourage removal of data items from their context which will tend to result in misapplication and misinterpretation. A report of expected carnation yield in a given soil mixture, for example, is nearly meaningless in the absence of additional contextual information: Where were the yields determined? What were the weather conditions? What technology was used? How were the yields measured? All these questions, and more, must be answered if the report on carnation yield is to be meaningful.

3. Avoidance of inconsistency in data is often cited as a major advantage of data base systems, and in many applications it is. In other situations, inconsistent data is an important characteristic of the entity being described, and removing inconsistency removes information. This is particularly true of biological data.

References

Anonymous. 1976. *Data base management.* DEC-00-XDBMA-B-D. Digital Equip. Corp. Maynard, MA, A brief discussion on file management and introduction to database management concepts.

Kroenke, D.M. 1977 *Data base processing.* Sci. Res. Assoc. Chicago, IL. An introductory text for computer professionals; the first chapter will be helpful to users.

Martin, James. 1980. *An end user's guide to data base.* Prentice-Hall, Inc. Englewood Cliffs, NJ. A useful and readable small book for the non-professional.

Martin, James. 1976. *Principles of data-base management.* Prentice-Hall, Inc. Englewood Cliffs, NJ. A professional text and a good advanced user's reference.

Dick Kingman, Executive Vice President
2785 N. Speer Blvd., Suite 230
Denver, Colorado 80211
Bulletin 417

NONPROFIT
ORGANIZATION
U.S. POSTAGE
PAID
Fort Collins, Colorado 80523
Permit Number 19

Direct inquiries to:
Office of the Editor
Horticulture Department
Colorado State University
Fort Collins, Colorado 80523