Woman C.P.A.

Volume 36 | Issue 4 Article 9

10-1974

Electronic Data Processing: Some Basic Concepts of Data Organization and Retrieval

Elise G. Jancura

Electronic Data Processing

Some Basic Concepts of Data Organization and Retrieval

Dr. Elise G. Jancura, CPA
The Cleveland State University
Cleveland, Ohio

Machine-readable data can take many physical forms. It can appear as holes in a punched card, as magnetic spots on a strip of tape, as magnetic characters on the bottom of a check. When read into the internal storage unit for processing, it will be converted into the internal storage format of that unit. Frequently, this data will be stored and processed in the decimal numbering system, but there are increasing numbers of applications where the information is stored in the binary numbering system because of its potential for efficient machine operation. With the developments in conversion techniques (both programmed techniques provided by the vendors and hardware techniques in the newer systems), the transition between the decimal and binary numbering systems is fairly convenient and of decreasing importance when choosing a numbering system. Of much more importance than the physical arrangement of the data is the logical organization and the meaningful relationships between the various pieces of information. Basically, these logical relationships hold regardless of the physical form of storage.

Basic Units of Information

Traditionally, a basic unit of information is a data record. A data record can be considered that collection of information which records an event, transaction, or happening which is to be recognized in the information system. For example, a sale of a given item to a particular customer would be a transaction which would be recorded in a unit of information called a record. This record will have certain subparts. As an example, information to be recorded about a sales transaction would include an identification of the customer and the amount of the sale. Each of these pieces of information about that sale would be recorded in a unit of information or a subpart of that record, called a field. Figure 1 is an illustration of a record, which is recorded on a punched card with five fields of information.

Sometimes one field takes on more importance than the rest for certain operational control purposes. For example, if this particular record, which is the information about a sales transaction, is to be used to update the accounts receivable file, then the field which contains the "customer number" becomes the field by which we recognize the relationship between this record recording the sale and the record within the accounts receivable master which contains the account balance of the customer involved. When a field is used to identify relationships between records, it is called a control field.

Fields are composed of a smaller unit of information called a character. A character, which is the smallest logical unit, can be related to the physical storage media in a sense that a character usually occupies a physical recording position. The amount of space that a piece of information takes within a physical storage media is a function of the number of characters in the logical record — usually one character occupies one recording position.

A physical input or output operation will frequently cause the transfer of one data record (also referred to as a logical record) between the input-output device and internal memory. However, there are some media or forms of physical storage in which the reading or writing of a single logical record with an input or output instruction is somewhat inefficient for the hardware. When this occurs, the logical record will sometimes be grouped with a number of other logical records into a unit called a block or physical record which can be read or written with the execution of a single input or output instruction. This technique of grouping these logical records so that they can be read several at

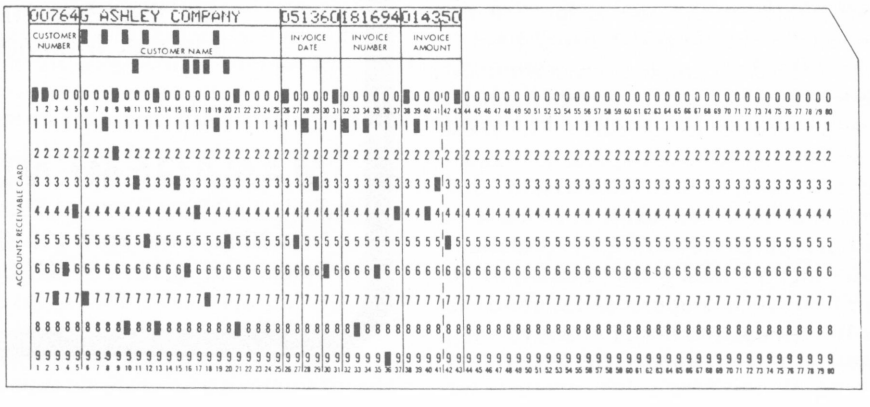

Figure 1. Data Fields Within a Record.

a time, or written several at a time, is called blocking. When this occurs, then one must distinguish between a logical data record and a physical data record. A logical data record is that group of information which records a single transaction or event. A physical record or block is that group of data characters which are read in or written out by the execution of a single input or output instruction.

A data file represents a collection of related records into a larger information unit. For example, all of the sales records for a given day (one record for each sale occurring in that day) could be collected into a file which would be the sales file for that day. Another example of a file would be the accounts receivable master file which is a collection of records showing the balance due from each customer.

Types of Data Files

Files basically are of two types. There are those files which could be designated as master files, which contain information that is usually considered permanent and which is periodically updated by transactions occurring in the normal operation of the organization. Sometimes these files are referred to as generation files. Once a file is updated, another version or another generation of that file exists. For example, the December 31 version of a master file will represent a different generation than the January 31 version which is produced by the use of January transactions to update the December 31 balances.

The frequency with which a master file is updated is dependent upon the processing cycle of the organization involved, but the principle is basically the same — master files are periodically updated; therefore, there are several generations of the same file. One of the concerns in an installation is to maintain proper identification not only of the file but also of the generations of the file.

Another type of file is that which might be characterized as a transaction file. Transaction files are not to be updated but instead contain data which will be used to update the master files. They are generally of a less permanent nature than the master files and are usually saved as long as they are needed for production of reports, or for use in an updating process, or for use in a reconstruction process.

The Trend Towards Integrated Files or Data Bases

Historically, as a given area was automated or a data processing application developed, the files were designed for that application, including the master and detail files necessary to accomplish the processing for that particular activity. As different areas were automated or separate applications developed, files for each of those applications were developed independently. As more and more of the processing activity and record-keeping activity of an organization was automated, many areas of duplication developed in the separate file systems for each application.

The duplication of data in several different sets of files represents several inefficiencies. First, it requires more storage when the same information is stored multiple times. Second, when a given transaction affects several different files which are not integrated or coordinated, it is necessary to use the transaction record in several different updating processes to update the different individual files. This can introduce inconsistencies between the similar records of separately updated files and the mere delay in execution of the multiple updating procedures can cause the information in the individual files to be at different stages of currency.

There has been a trend in the last few years to take these separate sets of files and to combine them into one integrated set of files referred to as a data base. An integrated file system provides greater processing efficiency by eliminating duplicate items currently existing in the separate sets of files. A transaction which would have affected several different sets of records in separate files through several processing runs can update the single integrated file in one processing pass.

The integrated data base, while providing greater efficiency, does, however, introduce a greater risk of damage to an organization's total information system. When each set of files is updated separately, an error in processing affects only one set of master files. The organization can continue to operate with all the rest of its data while reconstructing the particular file in error. In a single integrated file system, an error in the system can have more wide-reaching effect on the operational viability of the organization. Therefore, a great deal of effort must be expended in validating input data and additional controls procedures must be instituted to insure that valid transactions are properly processed against the correct file segments.

Data Organization and Retrieval

There are two basic approaches to data organization and retrieval — sequential and random. In a sequential file, the records are stored in some logical sequence. Usually that sequence is dictated by a field which identifies the record. This field is called the control field. For example, the records in a payroll file would be in sequence by employee number. Sequential organization has the advantage of simplicity and facilitates very efficient machine retrieval. When records are retrieved sequentially, physical movement within the input-output device is minimized. Sequential processing is characterized by passing the entire file and accessing each record in sequence. Therefore, it is most efficient in those instances where a large number of the records are active and would otherwise require access. In sequential retrieval all records in a given file are read in sequence. The efficiency of sequential retrieval is such that it is less time-consuming to access a few inactive records that would not otherwise have been read than to move the access mechanism of the input-output device in a non-sequential manner.

In those instances where many of the records in the file are relatively inactive, (that is, there are a large number of records but relatively few of them at any point need to be accessed) sequential organization and sequential retrieval loses some of its advantage. When there is a large enough percentage of inactive records in the file, the time advantage of reading the next sequential record is offset by the number of otherwise unnecessary read operations which are performed. In those circumstances a retrieval technique known as random retrieval is preferable. Random retrieval does not necessarily imply a lack of sequence in the records. Instead it refers to a particular processing technique in which only the active records are read. However, in certain processing environments, such as real-time systems, it is impossible to sequence the transactions. Under these circumstances it is necessary to use random retrieval when accessing the master file.

Random retrieval requires a hardware input-output device which is more sophisticated than that required by sequentially-organized records and which contains predetermined recording locations and an access mechanism which can be moved to those predetermined recording locations. Devices which have this ability are referred to as direct access devices (they have also been called random access devices). Direct access devices are represented in today's technology by the magnetic disk files, the drum files, and the data cells. Devices such as magnetic tape drives and the card readers are devices in which we have no predetermined recording positions other than the start of the file.

In simple unsophisticated systems there is frequently no difference between the logical organization of information as the user views it and the physical organization of the data as it is actually stored. The logical record represents that collection of information that is a meaningful unit to a processing program. Except for the concept of blocking (where several logical records may be grouped into one physical record), in most simple applications involving non-integrated files, the format of the logical record corresponds to the format of the physical record as it exists in the external storage device.

The movement toward more complex integrated collections of information that service multiple applications (a data base) has created some differences between the form of the logical record and the form of the actual physical record. The various input-output devices available with computer systems represent diverse physical characteristics just as the kinds of data and their uses vary. As information systems become more inclusive, a greater variety of processing needs are evolving at the same time that the advantages of a single integrated data base are causing a consolidation of data files.

In an integrated data base all of the individual pieces of information about a corporation's files (called segments or elements) are collected and physically stored and organized in such a way that any processing program can access any element within that data base. (This is in contrast to the approach in non-integrated files when each application sets up its own files and the file contains only records of a given type.) In the more complex environment of an integrated data base handling a variety of data and users, the need has developed for more versatile techniques for organizing and retrieving data. Frequently, the logical relationships between various data will no longer correspond directly to the physical structure of the data as stored in the external storage media.

In these integrated data systems, information storage has three basic levels of relationship: the user's concept of information involving fields, logical records and files; the system concept of data consisting of collections of stored records and files (frequently called data sets); and the device or storage concept of data consisting of physical blocks recorded on a storage media. Data management as shown in Figure 2 is the set of techniques by which the logical record of a user can be traced through the system organization to its physical storage location or by which a piece of information can be retrieved from a storage device to fulfill the need of a particular user for a given logical record of information.

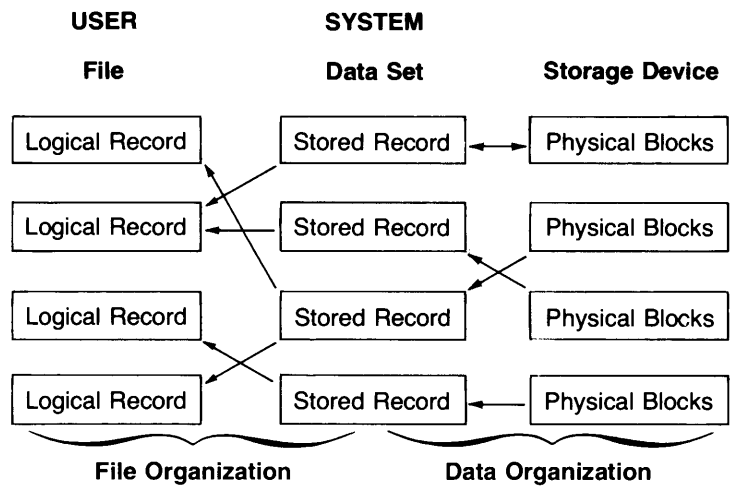

Figure 2. The Functions of Data Management

File organization represents that part of data management which expresses the set of relationships between the user's definition of logically related information and the system's collection of stored records or data sets. Data organization represents that part of data management which expresses the set of relationships between the system's data sets and the physical blocks of data as they are stored in a storage medium. Generally, the functions of data management (including techniques of file organization and data organization), are provided through vendor-provided programming support. In simple systems, data management is little more than blocking and deblocking routines. In complex systems, data management represents an extensive set of programmed routines for the identification, location, and retrieval of data.

Updating Techniques and Provision for File Reconstruction

File updating occurs when information from transactions is used to change the information in a master file, thereby creating a new generation of that master file. Basically, updating can be accomplished by one of two techniques — non-destructive updating or destructive updating. In a non-destructive updating environment, the old master file is kept intact. This is done by mounting the old file on a physically separate device from that device which will do the recording of the new file. This is the technique that is employed in magnetic tape processing. It allows for the retention of several generations of a file and it does, of course, facilitate reconstruction in those instances where necessary.

When a new generation has been created through an updating procedure, it becomes the input for the next processing cycle. The old generation is retained until completion of the next processing cycle to provide "back-up" in case the current generation is damaged. After two processing cycles there will be three generations of the master file: the most recent is often referred to as the "son"; the file generation used as input to the second processing cycle (and output from the first cycle) is referred to as the "father"; and the generation used as input to the first processing cycle is referred to as the "grandfather". (See Figure 3.) Once the "son" has been successfully created, the installation no longer needs to retain the "grandfather" generation or the transactions processed against it for reconstruction purposes. Of course, any reports or other printed record of the data represented by these files will still be available.

Reconstruction in computerized activities occurs when, for one reason or another, original data in their machine-readable format are destroyed and have to be recreated. If proper thought is given to forms for saving data in their machine-readable form, this reconstruction procedure can be greatly facilitated. If, however, when machine-readable data are destroyed, the installation has to go back to original non-machine-readable forms to recollect that information, the process can be time-consuming at best.

Generally, transaction records which affect a given version of the master file are saved until such time as that master file has been used as the input to another

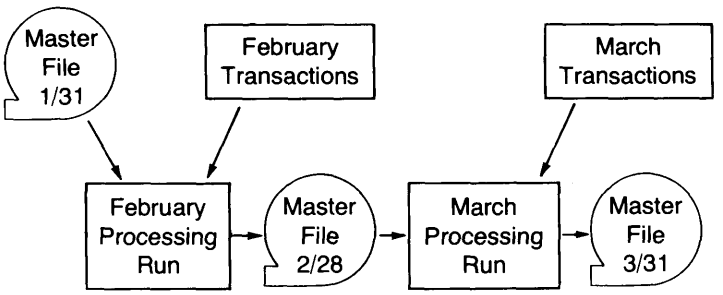

Figure 3. File Back-up with Generation Files

updating cycle in which a subsequent generation of the master file is produced. The procedure involves retention of several generations of the master file as well as all of the intervening transaction files.

Destructive updating occurs in those instances where the new version or the new generation of information is written on the same physical space as that previously occupied by the old version or old generation of the master record. This type of updating is frequently employed in direct access devices. Because the previous generation of data is destroyed in the process of updating, additional precautions must be taken in this approach to file updating. All possible checking of the transaction information should be performed before the data is used to assure its accuracy, and additional procedures should be performed during the updating procedure to insure that transactions are, in fact, being matched against the proper master transaction.

Since the "grandfather-father-son" retention technique is not viable in destructive updating as a reconstruction tool, some other procedure must be executed to accomplish this function. The usual approach is to make periodic copies of the master file, a process referred to as dumping the file. All transactions used in updating the master records since the last dump should be retained for reconstruction purposes until the next "dump" is made. Should any erroneous updating or other damage to the master file occur, it is possible to reconstruct the proper data by going back to the previous version and updating that with all intervening transactions. An alternate to periodic dumping of the entire file is the technique of logging changes to the master file as updating is being performed by writing the contents of the transaction and the master record before and after the update on a logging device (i.e., a reel of tape). Updating logs or periodic dumps can be made to another recording device similar to that containing the master file, they can be made to magnetic tape, or they can be printed out. The closer the form of the dump to that occupied by the master file, the faster the reconstruction procedure.

Summary

There is a great diversity of physical storage facilities, and file organization and retrieval techniques. Careful thought must be given to the choice of that combination of hardware and processing approach which will provide the most efficient and effective facility in relation to a given installation's circumstances. In addition to concerns for efficiency there is a continual need for adequate controls and protection of the data files and a procedure for reconstruction of those files should the need arise. This is especially true, given the trend towards centralized data bases and the frequently resulting operational dependency on the accuracy and continued availability of these information files.

Reviews

(Continued from page 22)

BEHAVIORAL ASPECTS OF ACCOUNTING, Michael Schiff and Arie Y. Lewin, editors; Prentice-Hall Inc., Englewood Cliffs, N.J., 1974; 408 pages, paperback.

Both undergraduate and graduate accounting curricula are currently including more emphasis on the behavioral sciences. A text developed for use in some newly designed courses may be of use and interest to accountants concerned with updating their skills.

The format is a series of 25 articles from economics, management science, and finance, divided into five categories: Theory of the Firm and Managerial Behavior, Budgeting and Planning, Decision Making, Control and Financial Reporting. Each section is preceded by a short introduction to the reading, including an overview of why each article was selected or what it should explain or demonstrate. A list of discussion questions follows each section. A "Selected Bibliography" of 8 pages concludes the book.

As always in a compilation of readings, some articles are very readable, others less so. Some are reprints from easily obtained journals, others would require use of a fairly well-stocked library. The convenience of the collection in one not-too-large paperback makes the volume more accessable to the accountant with minimum time.

A few of the articles contain functional symbols and equations, statistical techniques such as correlation coefficients, complex graphs, and flow charts. Others present their information totally in verbal form.

Most accountants have had some exposure to management, statistics, and economics. This paperback builds on and expands that knowledge, which perhaps has become out-of-date. The fields of management and economics have changed and expanded in recent years and have increasing implications for accounting today.

This reviewer suggests that the reader sample the book and read the introduction to each section. Benefits are not restricted to those who read the book from cover to cover.

M.E.D.