c. Data preparation, ie, transcription and verification.
d. Possible conversion from one medium (eg, diskette) to another (eg, magnetic tape cartridge).
e. Input of data to the computer for validation.
f. Sorting.
g. Control - all stages must be controlled.

Not all data will go through every stage and the sequence could vary in some applications. Even today, a high proportion of input data starts life in the form of a manually scribed or typewritten document and has to go through all the stages. However, efforts have been made to reduce the number of stages. Progress has been made in preparing the source document itself in a machine-sensible form so that it may be used as input to the computer without the need for transcription. In practice, the method and medium adopted will depend on factors such as cost, type of application, etc

Character recognition

The methods described so far have been concerned with turning data into a machine-sensible form as a prerequisite to input. By using Optical Character Recognition (OCR) and Magnetic Ink Character Recognition (MICR) techniques, the source documents *themselves* are prepared in a machine-sensible form and thus *eliminate* the transcription stage. Notice, however, that such characters can *also* be recognised by the human eye. We will first examine the devices used.

Document readers

Optical readers and documents. There are two basic methods of optical document reading:

a. Optical Character Recognition **(OCR).**

b. Optical Mark Recognition (OMR).

These two methods are often used in conjunction with one another, and have much in common. Their common and distinguishing features are covered in the next few paragraphs.

of an optical reader.

a. It has a document-feed hopper and several stackers, including a stacker for "rejected" documents.

b. Reading of documents prepared in optical characters or marks is accomplished as follows:

i. Characters. A scanning device recognises each character by the amount of reflected light (ie, OCR). The method of recognition, although essentially an electronic one, is similar in principle to matching photographic pictures with their negatives by holding the negative in front of the picture. The best match lets through the least light.

ii. Marks. A mark in a particular position on the document will trigger off a response. It is the *position* of the mark that is converted to a value by the reader (ie, OMR). The method involves directing thin beams of light onto the paper surface which are reflected into a light detector, unless the beam is absorbed by a dark pencil mark, ie, a mark is recognised by the reduction of reflected light.

Note. An older method of mark reading called mark sensing involved pencil marks conducting between two contacts and completing a circuit. c. Documents may be read at up to 10,000 A4 documents per hour.

Features of a document.

a. Documents are printed in a stylised form (by printers, etc, fitted with a special typeface) that can be recognised by a machine. The stylised print is also recognisable to the human eye. Printing must be on specified areas on the document.

b. Some documents incorporate optical marks. Predetermined positions on the document are given values. A mark is made in a specific position using a pencil and is read by the reader.

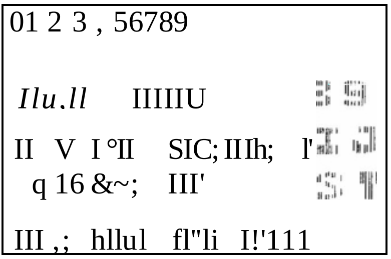
c. Good-quality printing and paper are vital.

d. Documents require to be undamaged for accurate reading. e. Sizes of documents, and scanning area, may be limited.

Magnetic ink reader and **documents.** The method of reading these documents is known as Magnetic Ink Character Recognition (MICR).

Features of magnetic ink readers.

a. Documents are passed through a strong magnetic field, causing the iron oxide in the ink encoded characters to become magnetised. Documents are then passed under a read head, where a current flows at a strength according to the size of the magnetised area (ie, characters are recognised by a magnetic pattern). b. Documents can be read at up to 2,400 per minute.



Features of documents.

a. The quality of printing needs to be very high.
b. The characters are printed in a highly distinctive type style using ink containing particles of iron oxide, which gives the required magnetic property. Examine a bank cheque for a further example.

Optical character recognition (OCR)

a. Technique explained.

i. Alphabetic and numeric characters are created in a particular type style, which can be "read" by special machines. The characters look so nearly like "normal" print that they can *also* be read by humans.

ii. Characters are *created* by a variety of machines (eg, line printers. typewriters, cash registers, etc) fitted with the special type face.

iii. The special optical character-reading machines *can* be linked to a computer. in

which case the data is read from the document into the processor.

b. Applications. OCR is used extensively in connection with billing, eg, gas and electricity bills and insurance premium renewals. In these applications the bills are prepared in OC by the computer, then sent out to the customers, who return them with payment cheques. The documents re-enter the computer system (via the OCC reader) as evidence of payment. This is an example of the "turnaround" technique. Notice that no transcription is required.

c. OCR/keyboard devices. These permit a combination of OCR reading with manual keying. Printed data (eg, account numbers) is read by OCR; hand-written data (eg, amounts) is keyed by the operator. This method is used in credit card systems.

Optical mark reading (OMR)

a. Technique explained. Mark reading is discussed here because it is often used in conjunction with OCR, although it must be pointed out that it is a technique in *itself.* Positions on a document are given certain values. These positions when "marked" with a pencil are interpreted by a machine. Notice it is the "position" that the machine interprets and that has a predetermined value.

b. Application. Meter reader documents are a good example of the use of OMR in conjunction with OCR. The computer prints out the document for each customer (containing name, address, *last* reading, etc,) in OC. The meter reader records the current reading in the form of "marks" on the same document. The document reenters the computer system (via a reader that reads OC *and OM)* and is processed (ie, results in a bill being sent to the customer). Note that this is another example of a "turnaround document".

Magnetic ink character recognition (MICR)

a. Techniques explained. Numeric characters are created in a highly stylised type by special encoding machines using magnetic ink. Documents encoded thus are "read" by special machines.

Week Five Lecture Note

Computer File Concepts

Files and file processing - part introduction

1. Files are named collections of stored data. Files tend to be too large to be held in main storage and are therefore held on backing storage devices such as magnetic disks. When data in a file is required for processing the file is read into main storage in manageable amounts.

2. Named programs are also often held on backing storage ready for use and may be regarded as "files" too. In this Part of the text we are primarily concerned with files stored for data processing which we may call "data files" to distinguish them from files containing programs. The term "data file" is interpreted broadly in this context to include text files.

3. Traditionally individual programs were written to process individual files or groups of files and this is still the case for many simple or specialist applications. However, files are also used as the building blocks for databases as will be described in later segments. As we will see, the database approach to the processing of data has some significant advantages over file processing in many situations but to understand why it is first necessary to have a basic grounding on how file processing works. Also, most of the basic file processing methods are still used within database systems although their use may not be evident to a user. This means that file processing is important both in its own right and as an underlying feature of database processing.

4. The examinations at which this text is aimed vary in the emphasis they give to file processing and in the extent to which they have shifted emphasis away from file processing to databases. The more detailed aspects of files organisation and access together with specific programming methods are not required for all examination courses so the reader would be wise to check syllabus details.

5. Others discuses the concepts of magnetic files and describes the methods of organising a file on *disk* (the main on-line storage medium for files) and how access is made to the records in such a file.

6. In the interests of clarity only one storage device (disk) is used as a basis for discussing computer files. More generally, the devices for file storage can be classified as follows: a. Disk (Magnetic hard or floppy, plus optical) - Direct Access Storage (DAS).
   Used for on-line file processing.
   b. Magnetic tape - Serial Access Storage (SAS). Once used extensively for on-line file processing but now mostly used to store files off-line e.g. for backup or archiving.

7. The *general* principles discussed with regard to *disk* can be applied to other SAS media.

Introduction

The purpose of this segment is to look at the general concepts that lie behind the subject of computer files before going on to discuss the different methods of organising them.

At all times the term "file" will refer to computer data files.

Purpose. A file holds data that is required for providing information. Some files are processed at regular intervals to provide this information (eg, payroll file) and others will hold data that is required at regular intervals (eg, a file containing prices of items).

There are two common ways of viewing files:

a. Logical files. A "logical file" is a file viewed in terms *of what* data items its records contain and *what* processing operations may be performed upon the file. The user of the file will normally adopt such a view.

b. Physical files. A "physical file" is a file viewed in terms of *how* the data is stored on a storage device such as a magnetic disk and *how* the processing operations are made possible.

A logical file can usually give rise to a number of alternative implementations. These alternatives are considered in later segments.

Elements of a computer file

A file consists of a number of records. Records were defined in 17.37. Here We consider records in a slightly different way because we are concerned with the way they are commonly stored. Each record is made up of a number of **fields** and each field consists of a number of **characters.**

Clock number	*Employee's name*	*Date of birth*	[redacted]	*Grade*	*Hourly rate*
1201	P J Johns	06 12 45	F	4	850

Field Field characters `P' `J' `J', etc

Notes of 1. Clock number is key field (15.10) 2. Grade is coded
3. Hourly rate is expressed in pence

a. Character. A character is the smallest element in a file and can be alphabetic, numeric or special.

b. Field. An item of data within *a record* is called a field - it is made up of a number of *characters*, eg, a name, a date, or an amount.

c. Record. A record is made up of a number of related fields, eg, a customer record, or an employee payroll record (see Fig. 19.1).

Alternative terminology

The terminologies of record, field and character are firmly established as a means of describing the characteristics of files in general situations. However, the use of this terminology can lead to excessive attention being directed towards physical details, such as how many characters there should be in a field. Such issues can divert attention from matters of high priority, such as what fields should be recorded in order to meet the information needs of the user. To overcome this difficulty, two alternative bets of terms have been developed, one set for physical files, and the other set for logical files. They are:

a. For physical files.
   i. Physical record.
   ii. Field.
   iii. Character (a physical feature).
b. For logical files.
   i. Logical record - an "entity".
   ii. Data item - "attributes" of the "entity".

Entities are things (eg, objects, people, events, etc.) about which there is a need to record data, eg, an item of stock, an employee, a financial transaction, etc. The individual properties of the entity, about which data is recorded, are its "attributes", eg, the attributes of an invoice (entity) will include the "name"; "address"; "customer order number"; "quantity"; "price"; "description".

A logical record is created for each entity occurrence and the logical record contains one data item for each occurrence of the entity's attributes, eg, the "customer's name" would be a data item and there would be one only in the logical record whereas the attribute "quantity" would have as many data items as there are entries on the invoice.

The relationship between the various terms used is summarised in Fig 19.2. (opposite)

Things about which there is a need to record data	Entities	Q each entity has a number of *attributes*
How the data is recorded	Logical records (1 per entity occurrence)	O each logical record contains a
Physical details of how the data is recorded	Physical record (1 or more per logical record)	71 each physical record contains a number of *fields*

Entities and Attributes.

Types of files

Master file. These are files of a fairly permanent nature, eg, customer ledger, payroll, inventory, etc. A feature to note is the regular *updating* of these files to show a current position. For example customer's orders will be processed, increasing the "balance owing" figure on a customer ledger record. It is seen therefore that master records will contain both data of a static nature, eg, a customer name and address, and data that, by its nature will change each time a transaction occurs, eg, the "balance" figure already mentioned.

b. Movement file. Also called transaction file. This is made up of the various transactions created from the source documents. In a sales ledger application the file will contain all the orders received at a particular time. This file will be used to update the *master file.* As soon as it has been used for this purpose it is no longer required. It will therefore have a very short life, because it will be replaced by a file containing the *next* batch of orders.

c. Reference **file.** A file with a reasonable amount of permanency. Examples of data used for reference purposes are price lists, tables of rates of pay, names and addresses.

Access to files

Key fields. When files of data are created one needs a means of access to particular records within those files. In *general* terms this is usually done by giving each record a "key" field by which the record will be recognised or identified. Such a key is normally *a unique identifier* of a record and is then called the **primary** key. Sometimes the primary key is made from the combination of two fields in which case it may be called a composite key k r compound key. Any other field used for the purpose of identifying records, or sets of records, is called a secondary key. Examples of primary key fields are:

a. Customer number in a customer ledger record.

b. Stock code number in a stock record.

c. Employee clock number in a payroll record.

Not only does the key field assist in accessing records but also the records themselves can, if required, be sorted into the sequence indicated by the key.

Storage devices

Mention is made here of the two storage devices that may be considered in connection with the storage of files (ie, physical files). a. Magnetic or optical disk. These are direct access media and are the *primary means of storing files on-line.*

b. Magnetic tape. This medium has significant limitations because it is a serial access medium and therefore is the *primary means of storing files off-line.*

These characteristics loom large in our considerations about files in the segments that follow. Note then that they are inherent in the *physical* make-up of the devices and will clearly influence what *types* of files can be stored on each one, and how the files can be *organised* and accessed.

Processing activities

We will need to have access to particular records in the files in order to process them. ;:

,. The major processing activities are given below:

a. Updating. When data on a master record is changed to reflect a current position, eg, updating a customer ledger record with new orders. Note that the old data on the record is replaced by the new data.

b. **Referencing.** When access is made to a particular record to ascertain what is contained therein, eg, reference is made to a "prices" file during an invoicing *run* Note that it does *not* involve any alterations to the record itself.

c. **File maintenance.** New records must be added to a file and records need to be deleted. Prices change, and the file must be altered. Customers' addresses also change and new addresses have to be inserted to bring the file up to date. These particular activities come under the heading of "maintaining" the file. File maintenance can be carried out as a separate run, but the insertions and deletions of records are sometimes *combined* with updating.

d. **File enquiry or** interrogation. This is similar in concept to referencing, it involves the need to ascertain a piece of information from, say, a master record. For example, a customer may query a statement sent to him. A "file enquiry" will get the data in dispute from the record so that the query may be settled.

Fixed-length and variable-length records

The question whether to use records of a fixed or variable length is one that usually does not have to be considered in manual systems.

a. **Fixed.** Every record in the file will be of the same fixed number of fields and characters and will never vary in size.

b. Variable. This means that not *all* records in the file will be of the same size. This could be for two reasons:

i. Some records could have more *fields* than others. In an invoicing application, for example (assuming a 6-character field to represent "total amount for each invoice"), we would add a new field to a customer record for each invoice. So a customer's record would vary in *size* according to the *number* of invoices he had been sent.

ii. Fields *themselves* could vary in size. A simple example is the "name and address" field because it varies widely in size.

It should be noted, however, that in the examples at a fixed-length record *could* be used. The record could be designed in the first instance to accommodate a fixed number of *possible* invoices. This means that records with less than the fixed number of invoices would contain blank fields. Similarly in the figure above, the field could be made large enough to accommodate the *largest* name and address. Again records with names and addresses of a smaller number of characters would contain blanks