COURSE CODE: CSC 315
COURSE TITLE: File Organisation and Data Processing
NUMBER OF Units: 3 Units
Course Duration: Three hours per week

COURSE DETAILS:
Course Lecturers: Dr O. Folorunso and Dr. (Mrs.) O. R. Vincent
Email: folorunsolusegun@yahoo.com, vincent.rebecca@gmail.com
Office Location: Room C and Room B201, COLNAS
Consultation hours: 12-2pm, Wednesdays & Fridays

Course Content: DATA PROCESSING-DATA PROCESSING CYCLE, DATA PROCESSING METHODS, COMPUTER MODES OF DATA PROCESSING, DATA VALIDATION TECHNIQUES- BATCH CONTROL, ON-LINE TRANSACTIONS CONTROL, THE DATA HIERARCHY, FILE PROCESSING

PHYSICAL STORAGE CONSIDERATION- WORSTATIONS AND TERMINALS,

FEATURES OF VDU TERMINALS, OUTPUT DEVICES,

DATA CAPTURE AND DATA ENTRY- FEATURES OF DATA ENTRY, PROBLEMS OF DATA ENTRY, OPTICAL CHARACTER RECOGNITION (OCR), OPTICAL MARK READING (OMR)

COMPUTER FILE CONCEPTS,FILE ORGANISATION AND ACCESS

Week One Lecture Note

1.0 DATA PROCESSING

This will discuss data and different validation techniques, for both on-line and batch systems of data entry. We shall also discuss different file accessing, organization and processing methods.

DEFINITION OF DATA

Data is the basic fact about an entity. It is unprocessed information. Examples are

a. Student records which contain items like the Student Name, Number, Age, Sex etc
b. Payroll data which contain employee number, name, department, date joined, basic salary and other allowances
c. Driving License which contain Driver's Name, Date of Birth, Home address, Class of license and its expiry date

Data can be regarded as the raw material from which information is produced by data processing. When data is suitably processed, results (or output data) are interpreted to

derive *information* that would assist in decision- making.

DATA STORAGE UNITS ON THE COMPUTER

Data is stored in the computer in a binary form. The units used to refer to this binary data are as follows:

Term	*Definition*
Bit	The smallest unit of data storage. A bit is either a 1 or a 0.
Nibble	4 bits This term is not commonly used.
i Byte	8 bits. The most commonly used storage unit.
Word	This term is architecture dependent. On some systems, a word is 16 bits; on others, a word is 32 or 64 bits.
Kilobyte (KB)	Even though *kilo* usually means 1,000, a kilobyte in computer terms is actually 1,024 bytes (because we like of 2).
Megabyte (MB)	The term megabyte denotes 1,024KB or 1,048,576 bytes.
Gigabyte (GB)	i A gigabyte is 1,024 megabytes or 1,073,741,824 bytes.
Terabyte (TB)	A terabyte is 1,024 gigabytes or 1,099,511,627,776 bytes.

DATA PROCESSING CYCLE

Data processing may be divided into five separate but related steps. They are

a. Origination
b. Input
c.Manipulatio
d. Output
e. Storage

Origination It should be kept in mind that "to process" means to do something with or to "manipulate" existing information so that the result is meaningful and can be used by the organization to make decisions. The existing information is generally original in nature and may be handwritten or typewritten. Original documents are commonly referred to as *source documents.* Examples of source documents are cheques, attendance sheets, sales orders, invoices, receipts etc. Producing such source documents, then, is the first step in the data processing cycle.

Input After source documents are originated or made available, the next step is to introduce the information they contain into the data processing system. This system may be manual, mechanical, electromechanical or electronic. However, our focus is on electronic data processing. This is done using any of the available input devices (keyboard, joystick etc).

Processing When input data is recorded and verified, they are ready to be processed. Processing or "Manipulation " involves the actual work performed on the source data to produce meaningful results. This may require performing any or all of the following functions - *classifying, sorting, calculating, recording and summarizing*

Output After input data has been fed into a data processing system and properly processed, the result is called output. Output can be either in summary or detail form. The

type of output desired must be planned in advance so that no waste of time and resources occur. Included with output is communication. Output is of little value unless it is communicated properly and effectively. The output is the ultimate goal of data processing. The system must be capable of producing results quickly, completely and accurately. The data processing cycle is incomplete without the concept of *control.* In an organization, controls depend basically upon the comparison of attained results with predetermined goals. When the results are compared, they either agree or differ. However, if a disagreement is detected, a decision is made to make. the necessary changes and the process repeated. This *feedback concept of control is* an essential part of data processing. That is, output is compared with a predetermined standard and a decision is made (if necessary) on a course of action, and is communicated to the stage where it is taken.

Storage Data related to or resulting from the previous four data processing steps can be stored, either temporarily or permanently for future reference and usage. It is necessary to store data, especially when it relates periodic reports, since they are used over and over again in other related applications.

A monthly attendance report or profit and loss statements will be useful in preparing annual reports. Stored information can either be raw, semi-processed or output data. Quite often, the output of one problem becomes the input to another. In the case of inventory, any unsold at the end of a year *(ending* inventory) become the *beginning* inventory for the next year. There are various ways of storing data, ranging from simple recording to storage in diskettes, hard disks, CDs etc.

DATA PROCESSING METHODS

Data originates in many different forms and they are many methods of processing:- manual, mechanical and electronic. The method used, however depends on its suitability to the task at hand. There are some that are best suited for electronic processing, while others are better done by manual methods.

Manual Method

This involves preparing data by means of such tools as pens, pencils, ledgers, files, folders etc. Improvements on these include using multi-copy forms, carbon paper etc. A good example is the daily marking of attendance register in school.

Advantages

a. They are generally cheap
b. Simple to operate
c. Easily adaptable to changes d. Easily accessible

Disadvantages

a. May take long time to complete
b. Cannot handle large of volume of work easily c. Generally prone to errors
d. Waste a lot of manpower

Mechanical Method

This method involves the use of a combination of manual processes and mechanical equipment to carry out the function. Examples are Typewriters, Calculators etc.

Advantages

a. Widely used in large and small organizations
b. Can serve as input to electronic system
c. Quality and level of output greatly improved as compared to manual method
d. Requires less manpower than the manual method

Disadvantages

a. Costly to purchase and maintain
b. Possibility of equipment breakdown
c. Produces lots of noise due to moving parts in the equipment
d. Usually slow in operation

Electronic Method

Here, the processing is done electronically by the system. There are two modes; batch processing and on-line processing.

Advantages

a. Faster analysis and results of processing
b. Handles complex calculations and problems
c. Can provide information in different and varied formats
d. Provides more accurate results than the other two methods
e. Work load capacity can be increased easily without hitches f. Provides for standardization of method g. Frees staff from clerical tasks for other tasks e.g. planning

Disadvantages

a. Initial acquisition cost may be high as well as maintenance costs b. Specialist personnel may be required
c. Decreased flexibility as tasks become standards

COMPUTER MODES OF DATA PROCESSING

There are two modes of computer data processing; **Batch Processing and On-line Processing.**

Batch Processing

A method of processing information in which transactions are accumulated and stored until a specified time when it is necessary or convenient to process them as a group. This method is usually adopted in payroll processing and sales ledger updates.

On-line Processing

A method of processing information in which, transactions are entered directly into the computer and processed immediately. The on-line method can take different forms. These forms are examined below.

Real Time Processing This is an on-line processing technique in which a transaction undergoes all the data processing stages immediately on data capture. This method is used in Airline ticket reservation and modern retail banking software.

Multiprogramming - This method permits multiple programs to share a computer system's resources at any one time through the concurrent use of the CPU. By concurrent use, we mean that only one program is actually using the CPU at any given moment, but that the

input/output needs of other programs can be serviced at the same time. Two or more programs are active at the same time, but they do not use the same computer resources simultaneously. With multiprogramming, a set of programs takes turns using the processor.

Multitasking - This refers to multiprogramming on single-user operating system such as those in microcomputers. One person can run two or more programs concurrently on a single computer. For example, the user can be working on a word-processing program and at the same time be doing a search on a database of clients. Instead of terminating the session with the word processing program, returning to the operating system, and then initiating a session with the database program, multitasking allows the display of both programs on the computer screen and allows the user to work with them at the same time.

Time Sharing - This capability allows many users to share computer-processing resources simultaneously. It differs from multiprogramming in that the CPU spends a fixed amount of time on one program before moving on to another. In a time-sharing environment, the different users are each allocated a tiny slice of computer time. In this time slot, each user is free to perform any required operations; at the end of the period, another user is given a time slice of the CPU. This arrangement permits many users to be connected to a CPU simultaneously, with each receiving only a tiny amount of CPU time. Time-sharing is also known as interactive processing. This enables many users to gain an on-line access to the CPU at the same time, while the CPU allocates time to each user, as if he is the only one using the computer.

Virtual Storage - Virtual storage was developed after some problems of multiprogramming became apparent. It handles programs more efficiently because the computer divides the programs into small fixed or variable length portions, storing only a small portion of the program in primary memory at one time, due to memory size constraints as compared program needs. Virtual storage breaks a program into a number of fixed-length portions called **pages** or variable length portions called **segments.** The programmer or the operating system determines the actual breakpoint between pages and segments. All other program pages are stored on a disk unit until they are ready for execution and then loaded into primary memory. Virtual storage has a number of advantages. First, primary storage is utilized more fully. Many more programs can be in primary storage because only one page of each program actually resides there. Secondly, programmers need not worry about the size of the primary storage area. With virtual storage, there is no limit to a program's storage requirements

Week Two Lecture Note

DATA VALIDATION TECHNIQUES

GIGO stands for Garbage-In, Garbage-Out. This means that whatever data you pass or enter

into the computer system is what would be processed. The computer is a machine and therefore has no means of knowing whether the data

supplied is the right one or not. To minimize such situations that may lead to the computer processing wrong data and producing erroneous output, data entered into a computer is validated within specific criteria to check for correctness before being processed by the system. This process is called **DATA VALIDATION.** We stated above that computer data processing is done in batch and **on-line processing modes** and we shall therefore discuss data validation techniques under each of these two modes.

Batch Control

This type of input control requires the counting of transactions or any selected quantity field in a batch of transactions prior to processing for comparison and reconciliation after processing. Also, all input forms should be clearly identified with the appropriate application name and transaction type (e.g. Deposits, Withdrawals etc). In addition, pre-numbered and pre-printed forms can be used where constant data are already printed and used to reduce data entry or recording errors.

Types of Batch Controls include:

- **Total Monetary Amount -** This is used to verify that the total monetary value of items processed equals the total monetary value of the batch documents.
- **Total Items -** This verifies that the total number of items included on each document in the batch agrees to the total number of items processed. For example, the total number of items in the batch must equal the total number of items processed.
- **Total Documents -** This verifies that the total number of documents in the batch equals the total number of documents processed. For example, the total number of invoices agrees with the number of invoices processed.
- **Hash Total -** Hashing is the process of assigning a value to represent some original data string. The value is known as hash total. Hashing provides an efficient method of checking the validity of data by removing the need for the system to compare the actual data, but instead allowing them to compare the value of the hash, known as the hash total, to determine if the data is same or different. For example, totals are obtained on an identifier (meaningless) data fields such as account number, part number or employee number. These totals have no significance other than for internal system control purposes. The hash total is entered at the start of the input process; after completion, the system re-calculates this hash total using the selected fields (e.g. account number) and compares the entered and calculated hash total. If the same, the batch is accepted or otherwise rejected.

On-Line Transactions Control

An advantage of on-line real time systems is that data editing and validation can be done up front, before any processing occurs. As each transaction is input and entered it can be operator prompted immediately an error is found and the system can be

designed to reject additional input until the error is corrected. The most important data edit and validation techniques are discussed below, but the list is by no means exhaustive.

- **Reasonableness Check -** Data must fall within certain limits set in advance or they will be rejected. For example, If an order transaction is for 20,000 units and normally not more than 100 units, then the transaction will be rejected.
- **Range Check -** Data must fall within a predetermined range of values. For example, if a human weighs more 150kg, the data would be rejected for further verification and authorization.
- **Existence Check -** Data are entered correctly and agree with valid predetermined criteria. For example, the computer compares input reference data like Product type to tables or master files to make sure the codes are valid.
- **Check Digit -** An extra reference number called *a check digit follows an* identification code and bears a mathematical relationship to the other digits. This extra digit is input with the data, recomputed by the computer and the result compared with the one entered.
- **Completeness Check -** A field should always contain data and riot zeros or blanks. A check of the field is performed to ensure that some form of data, not blanks or zeros is present. For example, employee number should not be left blank as it identifies that employee in the employee record.
- **Validity Check -** This is the programmed checking of data validity in accordance with predetermined criteria. For example, a gender field should contain only M(ale) or F(emale). Any other entry should be rejected.
- **Table Lookups -** Input data complies with predetermined criteria maintained in a computerized table of values. For example, a table maintains the code for each local government in the country and any number entered must correspond to codes found in the table.
- **Key Verification -** another individual using a program that compares the original entry to the repeated keyed input repeats the key-in process. For example, the account number, date and amount on a cheque is keyed in twice and compared to verify the keying process.
- **Duplicate Check -** New transactions are matched to those previously entered. For example, an invoice number is checked to ensure that it is not the same as those previously entered, so that payment is made twice.
- **Logical Relationship Check -** If a particular condition is true, then one or more additional conditions or data input relationship might be required to be true before the input can be considered valid. For example, an employee applying to be paid maternity leave allowance or employment date may be required to be at least eighteen years from date of birth and be a Male employee.

The Data Hierarchy

A computer system organizes data in a hierarchy that starts with bits and bytes and progresses to fields, records, files, and databases.
A **bit** represents the smallest unit of data a computer can handle. A group of bits, called a byte, represents a single character, which can be a letter, number or other symbol. A grouping of characters into a **word,** group of words or a complete number (e.g. a person's first name or age), is called a **field.** A group of related fields, such as a student's name, class, date admitted, age make up a record.
A group of records of the same type (e.g. the records of all students in the class) is

called a **file.** A group of related files (e.g. the personal history, examinations records and payments history files) make up a database.

A record describes an entity. An entity is a person, place, thing, or event on which we maintain information. An employee record is an entity in a personnel records file and maintains information on the employees in that organization. Each characteristic or quality describing a particular entity is called an **attribute.** For example, employee name, address, age, gender, date employed is an attribute each of the entity personnel. The specific values that these attributes can have can be found in the field of the record describing the entity.

Every record in the file contains at least one field that uniquely identifies that record so that the record can be retrieved, changed, modified or sorted. This identifier is called the key field. An example of a key field is the employee number for a personnel record containing employee data such as name, address, age, job title etc.

File Accessing Methods

Computer systems store files in secondary storage (e.g. hard disks) devices. The records can be arranged in several ways on the storage media, and the arrangement determines the manner in which the individual records can be accessed or retrieved.

Sequential Access File Organization - In sequential file organization, data records must be retrieved in the same physical sequence in which they are stored. Sequential file organization is the only method that can be used on magnetic tape.(e.g. data or audio tape). This method is used when large volumes of records are involved and it is suitable for batch -processing as it is slow.

Direct/Random Access File Organization - This is a method of storing records so that they accessed in any sequence without regard to their actual physical order on the storage media. This method permits data to be read from and written back to, the same location. The physical location of the record in the file can be computed from the record key and the

physical address of the first record in the file, using a *transform algorithm*, not an index. (The transform algorithm is a mathematical formula used to translate the key field directly into the record's physical location on disk.) Random access file organization is good for large files when the volume of transactions to be processed against the file is low. It is used to identify and update an individual's record on a real-time basis. It is fast and suitable for on-line processing where many searches for data are required. It is faster than sequential file access method. An example is an on-line hotel reservation system.

Index Sequential Access Method - This file access method directly accesses records organized sequentially using an index of key fields. An index to a file is similar to the index of a book, as it lists the key fields of each record and where that record is physically located in storage to ensure speedy location of that record. ISAM is employed in applications that require sequential processing of large numbers of records but occasionally require direct access of individual records. An example is in airline reservation systems where booking can be taking place in different parts of the world at the same time accessing information from one file. ISAM allows access to record in the most efficient manner.

Flat File - Supports a batch-processed file where each record contains the same type of data elements in the same order, with each data element needing the same number of storage spaces. Supports a few users' needs. It is inflexible to changes. It is used to enter data into an application automatically in a batch mode, instead of record by record. This process of automatic batch data entry is also referred to as a File Upload process.

Database File - A database supports multiple- users needs. The records are related to each other differently for each file structure. Removes the disadvantages of flat files.

Object Oriented File Access - Here, the application program accesses data objects and uses a separate method to translate to and from the physical format of the object.

File Processing

I

Different processes can be performed on files stored in the computer system. These processes include:

- **Updating -** The process of bringing information contained in the file up to date by feeding in current information
- **Sorting -** Arranging the records in a file in a particular order (e.g. in alphabetical or numerical order within a specified field)
- **Merging -** Appending or integrating two or more files into a bigger file
- **Blocking -** This is to logically arrange the records in a file into fixed or variable blocks or sets that can be treated as a single record at a time during processing. The gap between each block is known as the inter-block gap.

- **Searching -** This involves going through a whole file to locate a particular record or a set of records, using the key field.
- **Matching -** This involves going through a whole file to locate a particular record or a set of records, using one or a combination of the file attributes or fields.

Week Three Lecture Note

Physical storage consideration

"Volume" is a general term for any individual physical storage medium that can be written to or read from. Examples include: a fixed hard disk, a disk pack, a floppy disk, a CD-ROM, a disk cartridge or a tape cartridge.

Initialization. Before a disk may be recorded upon it normally has to be initialized which involves writing zeroes to every data byte on every track. A special program is normally supplied for this purpose. Clearly, the re-initialization of a disk effectively eliminates all trace of any existing data.

Formatting. In addition to Initialization the disk has to be formatted which means that a regular pattern of blank sectors is written onto the tracks. In the case of floppy disks the "formatting" program normally combines formatting with Initialization. On magnetic tapes the format is defined when the tape is mounted on the drive. Blocks of data are then formatted as they are written to the tape. The format determines the effective storage capacity of the volume. For example, a double sided floppy disk with 80 tracks per side and 9 sectors per track with each sector containing 512 data bytes will have a storage capacity of 720 Kbytes (ie, 9 x 40 x 2 x 512 bytes). Formats depend upon the manufacturer and operating system used. If data is to he transferred from one computer to another not only must the volume be physically interchangeable between drives the volume format must be compatible