NATIONAL OPEN UNIVERSITY OF NIGERIA

SCHOOL OF SCIENCE AND TECHNOLOGY

COURSE CODE: CIT 141

COURSE TITLE: INFORMATION STORAGE AND RETRIEVAL I

COURSE
GUIDE

CIT 141
INFORMATION STORAGE AND RETRIEVAL I

Course Team Dr. F. A. Ehikhamenor (Course Developer/Writer)

NATIONAL OPEN UNIVERSITY OF NIGERIA

National Open University of Nigeria
Headquarters
14/16 Ahmadu Bello Way
Victoria Island, Lagos

Abuja Office
5 Dar es Salaam Street
Off Aminu Kano Crescent
Wuse II, Abuja

e-mail: centralinfo@nou.edu.ng
URL: www.nou.edu.ng

Published by:
National Open University of Nigeria

Printed 2013

ISBN: 978-058-352-1

CONTENTS **PAGE**

INTRODUCTION

Information Storage and Retrieval I (CIT 141) is a one-semester course in the first year of the Bachelor of Science degree in Communications Technology, Computer Science, and Data Management. It is an introductory course that will be followed in the second year by the more advanced Information Storage and Retrieval II. As an introductory course, CIT 141 is simple and non-technical.

The course consists of 18 units of reading material, which should engage you for about 17 weeks. Each unit is designed to provide reading material for two to three hours of study. You have no additional burden of reading supplementary materials. This course does not require prior skills in any area of knowledge other than the general admission requirements. The approach at this level is to help the student to appreciate the basic characteristics of information, the need for good organisation of information, and the fundamental concepts of storage and retrieval. You are expected to acquire some computer skills by the time you complete this course so that you will be able to work through the more advanced course in your second year.

The rest of this guide will tell you what you are expected to learn in this course, how to work through the course, the content of the course, useful information on exercises and assignments, and how to get the most out of the course.

WHAT YOU WILL LEARN IN THE COURSE

Information has become a crucial commodity that has a market value. It has become a major resource to be acquired, managed and even controlled for corporate and national interests. In recent decades, much attention has been given to the complementary issues of storage and retrieval of information.

In this course, you will begin to appreciate the general concerns about the importance of information and the need for efficient systems to process, store and retrieve it. You will also be introduced to the basic technologies necessary for these tasks.

COURSE AIMS

The aims of this course are to help you:

- master the techniques of preparing data and information for storage

- learn to store data and information in appropriate formats and in suitable media
- acquire skills in retrieving data and information from an information system.

COURSE OBJECTIVES

In order to achieve the aims of this course, a number of objectives are specified for each unit besides the following overall objectives. By the time you complete this course, you should be able to:

- explain the concepts of data, data processing, and information
- distinguish between document and information and describe the processes of documentation
- classify information into subject categories
- analyse data and information for the purpose of assigning it to appropriate subject classes
- correctly assign key words to be used for retrieval purposes
- describe the basic architecture of a computer and the role of computers in storage and retrieval
- describe the storage media in use and the basic structure of records, files, and databases
- explain the functions of Database Management System
- explain the concepts of information retrieval
- discuss user characteristics and user needs, which are fundamental to information storage and retrieval
- correctly analyse requests for information and formulate search strategies
- retrieve information from an information system
- search the Internet
- evaluate the retrieval performance of an information system.

WORKING THROUGH THIS COURSE

To go through this course you are required to read the study units, answer the self-assessment exercises and do the assignment in each unit. The self-assignment exercises are meant to help you to reinforce what you have learnt. It will be very helpful to you to try and answer the questions first before looking at the answers. At the end of the unit, you will find an assignment, which will be marked by your tutor. Work diligently on it and submit your work to your tutor for grading. The tutor-marked assignments will constitute 30% of the total marks of the examination in this course.

You will need to have access to a computer and be familiar with the basic elements of a computer system. In due course, you will have practical exercises in the library and on the Internet, and so, you need to have access to these facilities.

It is expected that you will spend on the average two to three hours to study one unit and about 17 weeks to complete the whole course. However, you should realise that you are actually to work at your own pace. Below are the components of this course.

COURSE MATERIALS

1. Course Guide
2. Study Units
3. Self-Assessment Exercises
4. Tutor-Marked Assignments

STUDY UNITS

There are 18 study units which you will work through in this course. There are as follows:

Module 1

Unit 1	Data and Information
Unit 2	Document and Documentation Classification
Unit 3	Classification Unit
4	Subject Indexing
Unit 5	Indexing Language

Module 2

Unit 1	Computers in Information Storage and Retrieval
Unit 2	Storage Media
Unit 3	Records and Files
Unit 4	Databases
Unit 5	Database Management System

Module 3

Unit 1	Concepts of Information Retrieval
Unit 2	Role of Information Centres
Unit 3	Users and User Needs Information
Unit 4	Searching for Information
Unit 5	Document-Term Matrix

Module 4

Unit 1	Retrieval from the Internet
Unit 2	Evaluation of an Information Systems and Services: Part I
Unit 3	Evaluation of an Information Systems and Services: Part II

Units 1 to 5 of module 1 address the crucial requirements of understanding the characteristics of information and systematic organisation

Units 1 to 4 of module 2 take up the requirements for information processing and storage; while modules 3 and 4 are devoted to the principles of retrieval and evaluation of retrieval performance.

SELF-ASSESSMENT EXERCISES

These are embedded in the text of the study units. You should be able to answer the questions if you study the sections of the units very well.

TUTOR-MARKED ASSIGNMENTS

The tutor-marked assignments will be supplied to you with the units. It is absolutely necessary that you do the assignments and submit your work to your tutor.

ASSESSMENT

There are two aspects of the assessment of this course. The first aspect is the continuous assessment through the tutor-marked assignment. The second aspect is the final examination. The tutor-marked assignments will constitute 30% of the total marks of the examination in this course. The final examination will come at the end of the course. It will be a written examination that reflects the exact content of the course. The questions will not be different from the types you would have already been familiar with in the self-assessment exercises and tutor-marked assignment. The written examination will carry 70% marks.

HOW TO GET THE MOST FROM THIS COURSE

In this programme, you will not be sitting before any lecturer to receive lectures. The study units will replace the lecturer; you will be reading the study units instead of listening to a lecturer. You have the flexibility of being able to work through specially designed course materials at your own pace. You can also choose your time and place of study. The contents of the units will give you all the information and direction you need.

The units follow the same format. Each unit begins with a table of contents, which tells you at a glance what is covered in the unit. This is followed by an introduction to the subject matter of the unit and the relationship of the unit to the previous unit. Then follows the objectives in which you are told what you should be able to do by the time you complete the unit. It is advised that you use these objectives to guide your study. After the objectives, you come to the main body of the study unit. The text of the reading is presented in a simple direct style to engage your attention and assist your concentration. You are to go through the unit, section after section. Make sure you fully understand a section and that you have done the self-assessment exercise there before going to the next one. The conclusion that follows the main body of the unit gives you an overview of what you would have achieved in the unit. You should also refer to the objectives of the unit to assure yourself that they have been met. If you are not satisfied that you have achieved all that you were expected to achieve, just go through the unit again. The summary of the unit relates what you have learnt in the unit to the subject matter of the next unit, thus building a "bridge" between the two units. In this way you can see a logical connection between all the units.

You will find this course quite interesting and the study units quite readable. The only problem that you need to worry about is your ability to create a conducive environment for your study. You have to work out your timetable, time and place of study; and demonstrate a serious commitment to your study.

SUMMARY

This course should equip you with basic skills in information storage and retrieval. It is neither abstract nor highly theoretical. Both the course aims and objectives have been set out at the beginning of the guide.

They are all realisable and you should not have any problem realising them. It is hoped that you will find the course interesting and challenging and that you will enjoy reading the course material. We wish you brilliant success.

MAIN COURSE

CONTENTS PAGE

MODULE 1

UNIT 1 DATA AND INFORMATION

CONTENTS

1.0 INTRODUCTION

You will find out that many people use the terms "data" and "information" in such a way that you would think they mean the same thing. In this unit, you will learn to use the terms correctly; and also to appreciate the value of data and information.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain the meaning of data
- explain the concept of data processing
- explain the meaning of information.

3.0 MAIN CONTENT

3.1 Data

The word "data" is the plural form of the word "datum". Data may be regarded as symbols or figures that have potential value or to which meaning can be given. Now, let us consider how people record events. Think of the old man in the village who made nine strokes of chalk on the lintel of his front door to remind him that it was exactly nine hundred naira that he borrowed. Each time he paid back one hundred naira, he wiped away one stroke, when he was able to pay back a multiple of one hundred naira, he cleaned off the corresponding number of strokes. Would you not consider such a man as well organised? He kept accurate data. He could always tell how much of his debt was outstanding by counting the number of strokes left. Those strokes might not mean anything to someone else but they mean a lot for the old man. Look at the table below and try and make some meaning out of it.

Table 1: Three Ways of Recording Data

	Column 1	**Column 2**	**Column 3**
Sunday	●●●	𝍷𝍷𝍷𝍷	12
Monday	●●●●●	𝍸 𝍷𝍷𝍷𝍷	3
Tuesday	●●●●●	𝍷𝍷𝍷𝍷 𝍷𝍷𝍷	3
Wednesday	●●●●●●●	𝍸 𝍷	3
Thursday	●●●●●	𝍸	5
Friday	●●●	𝍸	6
Saturday	●●●●●	𝍷𝍷𝍷	7

You may not know the events that were recorded but you can see the frequency of occurrence for each day of the week. Column 1 shows how a little boy recorded the number of cars that came to his father's house during a particular week. Column 2 shows the number of cows a dealer sold in a space of one week. The record was kept by his son. Column 3 illustrates the number of phone calls a young lady received during a particular week.

If you ask all the pupils in a class in a primary school to write their names on a sheet of paper and against their names write their ages, you will have another set of data, namely age distribution of the pupils in the class. The statistical data of the last local government election in 12 wards of a local government is captured in Table 2.

Table 2: Data from a Local Government Election

Ward	Number of Registered Voter	Number of Votes
1	1,567	1,323
2	893	892
3	1,083	995
3	1,002	857
5	1,803	1,635
6	1,337	999
7	1,291	1,295
9	2,002	1,328
10	1,066	993
11	1,153	1,039
12	977	936

Now consider the votes on four motions in a state house of assembly.

The following data were generated.

Table 3: Data from the Votes on Four Motions in a State House of Assembly

	YES	NO	ABSTENTION
Motion 1	18	7	2
Motion 2	11	12	3
Motion 3	13	8	6
Motion 4	9	17	1

SELF-ASSESSMENT EXERCISE

What do you understand by the term "data"?

3.1.1 Data Processing

To put it simply, data processing is what we do to data in order to make some sense out of them. There are many ways of processing data. We may just inspect data and be able to see a pattern in them. From such a pattern we can make a statement about what has happened and even take a decision. This is visual inspection. Let us apply it to the sets of data in Table 1. The data in column I show that more cars came to the house on Wednesday. Column 2 not only tells us that more cows were sold on Monday, but that sales declined during the rest of the week. From column 3 we can say that the young lady received the highest number of phone calls on Sunday. We can add that she is more likely to receive more calls during the weekend.

We can build much story on the data in Table 2. In each ward, we can compare the number of registered voters with the number of those that voted. Naturally, not everyone who registered could have, voted. We can compute the voting rate for each ward as the percentage of the registered voters who voted. We can find the ward with the highest voting rate and the one that comes next. We can also rank the wards according to the number of registered voters and on the basis of the ranking predict in which wards to expert the highest number of voters. Now, here is a sticky point: what do you say when the number of voters is higher than the number who registered? A big mistake somewhere or what?

Now let us turn to Table 3. Can you say something about the popularity of the four motions? Obviously, motion 1 was the most popular. Motion 2 must have been highly controversial. Motion 4 was highly unpopular.

Besides visual inspection, rearrangement or sorting, data processing could take the form of arithmetic processes such as computing the sum or mean. Statistical techniques may also be used to show how much the data deviate from some reference value or the relationships within the data. Then we are able to make statements about the data as well as generalise our observations to other similar situations. Suppose we found that out of 420 boys that sat for mathematics in SSSE, 303 passed with credit and above; whereas out 392 girls who sat for it the number who passed with credit and above was 121. Then we can at least say that boys do much better than girls in mathematics in that school. If we collected our data from a number of schools across the country and got the same pattern, then we can say that in Nigeria, boys are better in mathematics than girls provided that the data were generated properly.

It is difficult to estimate the volume of information a person handles everyday. Unless a person is sleeping, his or her brain is always busy processing data and handling information. The brain receives both data and information through the eyes, for instance from what the person reads or sees; through the ears, for instance from what the person hears or listens to; through the nose, for instance from what the person smells, and through the senses of touch, exposure, physical contact and taste.

SELF-ASSESSMENT EXERCISE

Name four ways of processing data.

3.1.2 Information

In the paragraphs above, we considered the processing of data. We found that having done such processing, we were able to make some statements on the situations or events on which the data were obtained. Such statements could be used to guide our future response or action. This is the function of information: to guide a person on what to do, how and when to do it. Consequently, information may be defined as a fact or set of facts that can influence a person's response in a given situation. Ideally, information reduces or even eliminates a person's sense of uncertainty. We could also define information as the outcome of data processing.

Information is a vital resource. It would be impossible to live a normal life without having adequate information. Just imagine waking up one morning to find that you are alone in the house. Not even one other person is around in your house, in your compound, and in the neighbourhood. A thousand questions are racing through your mind, but there is not a soul around to answer you. You decide to move down some distance but still there is not a single person anywhere. What would you do? For how long would you be able to endure that experience?

In order to embark on tertiary education, you needed information to decide what programme to choose. After completing your application for admission, you waited with some anxiety for, information on whether you have been given admission or not. You did not know what else to do until you got that information. A businessman would need up-to-date information on the market situation, and changes in government policies that may affect his business. He would normally consult on a regular basis with his colleagues and share information with them. Today, people are realising more and more the value of information and so a science has developed around it. Information science is the science that deals with the generation, acquisition, organisation, storage, retrieval, dissemination and use of information, its characteristics as well as its impact at the individual, corporate and societal level.

Information plays a key role in every sphere of life. Information is at the core of success of both individuals and corporate bodies in commercial and business enterprises. Information confers competitive advantage on those who have it against their counterparts who do not have it. Information gives power. The countries that have the capacity to acquire or generate and manage information use it to improve their socio-economic status and advance ahead of other nations that do not have such capacity.

Countries in the Third World do not seem to fully appreciate the value of information. In many of these countries, most people in the civil service and in government seem to regard information as what government wants the citizenry to hear, the kind of releases made by the minister or ministry of information, or the news one hears on radio and television or reads in the newspaper. That is information quite alright; but it is "soft" information. The kind of information that confers power in our age, include, scientific, technological, economic and developmental information.

The crucial importance of information has also dictated that both organisations and national governments take appropriate measures to put in place the infrastructure necessary for managing and possibly for controlling it. More and more investment is being made in the establishment of information systems. National, regional and global computer and telecommunication networks have been developed for the management and communication of information. The countries in the Third World have been talking of a new information order. That is a reaction against the domination by the Western and industrialised countries of the global information industry.

3.2 Value of Information

The value of information is enhanced by its accuracy, relevance, timeliness, source, up-to-dateness, as well as the packaging format. If you have any reason to doubt the accuracy of a piece of information, you would not like to act on it. As a matter of fact, inaccurate information could be more disastrous than no information. If it becomes necessary to verify information coming from a particular channel, the cost of such information will be higher, and that may discourage the use of it.

It is quite, important that information be relevant to the purpose for which it is needed. If you would like to read something about the industrial revolution and a library staff gives you a book on the colonisation of Africa, would you be pleased? You may accept the book if you think you might have time to read it, but your need for information on the industrial revolution has not yet been satisfied. You may in fact reject the book. The book is still important and will be useful to someone else but not to you at that point in time.

The other consideration is timeliness. For information to be useful, it must be received in good time to make a difference in what the person who receives it is actually doing. If you wanted the information on industrial revolution in the course of preparing for an examination and you could not get it until after that examination you would no longer attach much importance to it.

The source of information could be very important. The source could have a high level of credibility so that the person using the information could do so with much confidence. Otherwise it could have some qualified level of credibility. Certainly you would be more assured of the authenticity of information when you hear that it has come from an impeccable source.

It must be noted that information has to be presented in a way that makes it easy to use. A brief summary may be enough and much better for a business executive than a voluminous report. Graphic representation may carry more impression than pages of a statistical report. A video presentation could be better appreciated than a written version on a particular subject. The reverse may be the case in another situation.

SELF-ASSESSMENT EXERCISE

What are the functions of information? Name six factors that determine the value of information.

4.0 CONCLUSION

In this unit you were introduced to the basic concepts of data and information. You should now be able to explain the meaning of data, data processing and information. You can now better appreciate the value of information and the factors that determine the usefulness of information.

5.0 SUMMARY

In this unit, you have learnt what data is and a number of ways of handling data in what is referred to as data processing. You also learnt what information is, the importance of information, and the factors that dictate the usefulness of information. In the next unit, you will learn to distinguish between information and document and you will study the processes of documentation.

6.0 TUTOR-MARKED ASSIGNMENT

Write an essay on "Data and Information". Your write-up should be between six and eight pages of typed A4 double-spaced, 12 points Times Roman. You should include the following:

i. Definition of data
ii. Ways of recording
iii. Data processing
iv. Definition of information
v. Importance of information
vi. Government and corporate roles in information management.

7.0 REFERENCE/FURTHER READING

Susan, A. (1972). *An Introduction to Computers in Information Science*. Metuchen, N.J.: Scarecrow.

UNIT 2 DOCUMENT AND DOCUMENTATION

CONTENTS

1.0 INTRODUCTION

In the last unit you learnt the basic concepts of data, data processing, and information. Now you are about to learn the difference between information and document and to be introduced into the whole process of documentation. You will see that documentation is the essence of research.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- distinguish between information and document
- describe the characteristics of documents
- explain the principles and history of documentation
- explain how research contributes to the growth of information and creation of knowledge
- explain the concept of bibliographic control.

3.0 MAIN CONTENT

3.1 Information and Document

A document is a source of information, and it is as useful as the information in it. A letter from your, uncle is a document. So also is every other letter you receive. Your birthday certificate is a document. It says something about your birth. Every book you buy and add to your collection is a document. The minutes of the meetings of your club are documents. Your photo albums and your photographs are more examples of document. Your diary is another document. We can add to the list such documents as an issue of a newspaper or a magazine, an audio tape with music or speech, a video recording of any event, a film strip of a local festival, directories, sales promotion brochure and leaflets, painting and maps. Now you can see that when we use the term "document" we are actually referring to a wide range of information sources. A document is therefore the vehicle conveying information.

SELF-ASSESSMENT EXERCISE

Examine the following list and identify the items that are not documents.

- Identity card
- Course guide
- Unused note book
- Video of "Things Fall Apart"
- Uncompleted application form
- Map of Nigeria.

3.2 Characteristics of Documents

We shall now examine some of the characteristic of a document. The first fact to note is that every document must originate from somewhere. In many cases, there is one individual or more persons who created the document. A person who created a document is the author. If the document is a book, the author did all the work of writing it. This course material you are reading now was written by somebody who is the author. A book may have one author or multiple authors. Sometimes, a number of authors contribute sections or chapters of a book while one of them or someone else compiles these contributions and edits them to produce the book. The latter is the editor. On the front back of such a book you will find something like

Democracy in Nigeria

Edited by O.O. Galu

Sometimes the originator of a document may not be just an individual or some individuals but rather an organisation, for example the National Universities Commission, the Ministry of Information, UNESCO, and so forth. That is a case of co-operative authorship. Even though somebody in the organisation wrote the material, it was written in the name of the organisation. Later in this course you will see the importance of author in locating information.

Another feature, of a document is the place of publication. When a document has been produced in a large quantity for sale or distribution, it is said to be published. The place where the work of preparing the document was done is the place publication. Usually, the place of publication is the place where the publisher has its main office. However, a publisher may publish a document in more than one place. Take any book and inspect the title page or the reverse of the title page and you will find where the book was published. There you will also find the publisher. There are a number of well-known publishers in Nigeria. There are also many little-known publishers.

After the author, the title of the document follows in a bibliographic record. The title is the name that the author gives to the document. The title could be in two parts: the main title followed by a sub-title, for example: "Computer Concepts: A User Perspective." Besides identifying a document, the title could also give a glimpse into what the document is about. The title above suggests that the book is about computers. It must be noted that this is not always the case. If you see a document with the title, "An Enterprise in Futility" would you be able to guess what it is all about?

Another feature that may identify a document is the edition. A new edition comes out when a document is revised. Revision is usually necessary for the purpose of correcting errors in a document, enhancing aspects of the document, supplying new information, expanding the scope of coverage and so forth. Some documents go through many revisions, which generate successive editions. Two different editions of the same document are actually regarded as two different documents.

The date of publication is no less important. The date of publication says how recent the document is. In some subject field publications become quickly out of date.

Lastly, we should mention the unique identification number that every book should have. It is called the ISBN (International Standard Book Number). It is a 13 digit number, for instance 978-1-846-14792-0. ISBN not only identifies the book, but also the publisher and the country of publication. Serial publications such as journals and magazines also have a unique identification number called ISSN (International Standard Serial Number).

You may wonder if every document has all the features described above. Not every document has them. A document that has not been published cannot have a publisher or place and date of publication. We may classify information source as published or unpublished. The value of an information source does not depend as much on whether or not it has been published as to the novelty and potential usefulness of the information it contains.

SELF-ASSESSMENT EXERCISE

- Name seven particulars of documents.

3.3 Documentation

Here is a question to thrash out, "How do documents come into existence?" Earlier we said that they are originated by someone or an organisation, especially in the case of textual documents. In the mind of anyone who creates document, the intention is to have a more-or-less permanent record of an event or phenomenon or ideas.

A major requirement for documentation is a language. People express their thoughts and expressions in a language they have learnt in whatever way. The language serves also as a medium for documenting their thoughts and expressions. Every ethnic community evolves its language and culture, which influence the worldview of its members. In other words, a language serves for the purpose of communication and documenting information. For instance, the language helps to typify an observation and assign it to an appropriate category.

In a country like Nigeria where oral culture still has a dominant role, documentation of oral information is very important. Festivals and cultural shows are better captured by video techniques. Efforts to reach the rural dwellers in Nigeria with information will be largely unsuccessful unless the information content is transmitted in a way that is compatible with their oral culture.

While we recognise the information needs of rural dwellers, we are going to pay more attention in this course to the information issues that are more pertinent to the literate segment of the society, especially because of the urgency of their needs and expected impact on national issues. A greater proportion of the information that will be stored or retrieved for their benefit is in textual form. We should begin by reviewing the historical development of documentation in textual form.

SELF-ASSESSMENT EXERCISE

What is the most important requirement for documentation?

3.4 History of Documentation

The earliest form of documentation was found in caves; and the format was pictorial. Then came the cuneiform and later the hieroglyphics of the Sumerians/Assyrians and Egyptians respectively. Their symbols as well as those of the Chinese were basically pictograms. Much later, the use of standardised characters was introduced in various parts of the world. This allowed the early civilisations to benefit from a well-developed means of literary communication. Documentation was done on various materials, which are unknown to our generation, including papyrus, codex, parchment, and velum.

The development of the book was boosted by the invention of the printing press, which occurred between 1353 and 1355. In the early years, only the Bible was printed and distributed in printed form. Later, the transactions of learned societies were printed. Other publications followed and the growth of publications took a dramatic turn, leading to the information explosion of the second half of the twentieth century. Beside the rapid growth of publications, the printing press facilitated the availability of documents, which in turn encouraged education, increase in literacy level, and a reading culture. Furthermore, libraries increased in number and their function both as custodians of documents and intermediaries to users grew in significance.

The printing press played a crucial role in documentation and remained unchallenged for several centuries. By mid-19th century advances in optical technology ushered in photography, which, in documentation, became a complement to the printing press. Photography was followed by radio transmission and telephony. Alongside these developments was the success in sound recording in phono discs. One technology became a springboard for another. So, before the end of the 19th century motion pictures became possible and cinema houses sprang up. From the 1930s, television began broadcasting news and educational and entertainment programmes to millions of homes. Finally, the computer came out as the central piece in information processing, storage and dissemination. Now computer networks have become the primary vehicle for conveying information all over the world.

SELF-ASSESSMENT EXERCISE

Name the most significant inventions that have facilitated documentation.

3.5 Creation of Knowledge

A major contribution to the information explosion phenomenon has come from research and scholarly communication. A scholar always wants to engage in research and to publish the findings of the research. He has to document the output of his research to establish his claim to his findings as well as to communicate or make them available to other scholars in a form that can be preserved for future generations. The cycle of activities that generate the information to publish is presented in Figure 1.

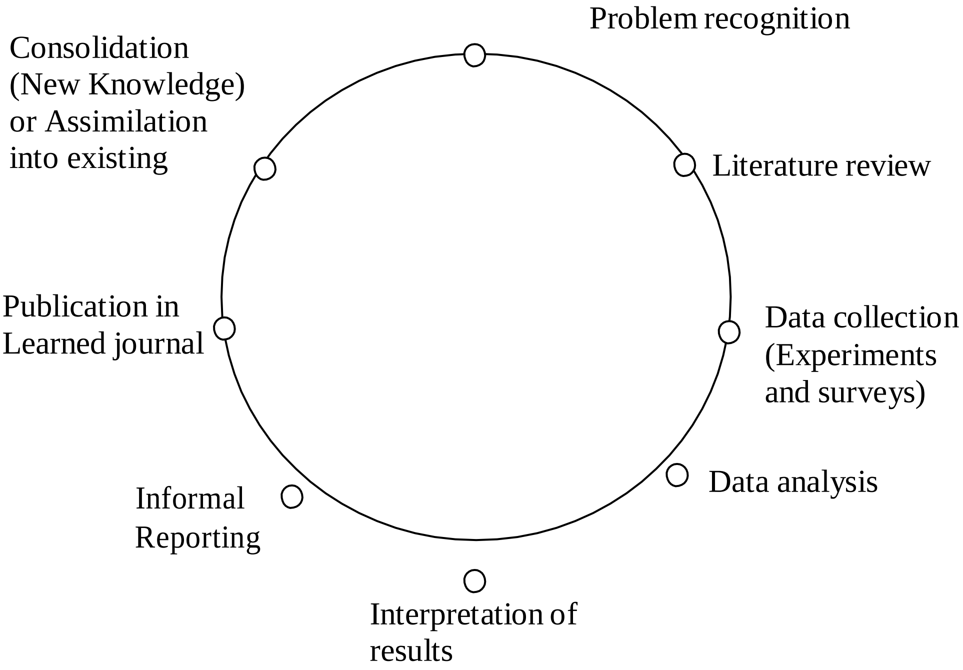

Problem recognition
Literature review
Data collection (Experiments and surveys)
Data analysis
Interpretation of results
Informal Reporting
Publication in Learned journal
Consolidation (New Knowledge) or Assimilation into existing

Fig. 1: Information Generation Cycle

There is the tendency to publicise the research findings in various informal outlets such as reports and conference proceedings; but ultimately the findings are published in a referred journal. This is the preferred outlet of communicating research results. Such outlets are called primary sources of information because they are the very first media that report new information coming directly from research efforts. Having being published, the research findings are subjected to years of use, critical assessment, and incorporation into existing body of knowledge and possibly to extend the frontiers of knowledge.

After the primary information sources are the secondary sources that draw the attention of information users to the primary publications. They include indexes and abstracts, bibliographies, guides, and so forth. The tertiary sources, for example textbooks, compile and integrate the scattered information in the primary and (sources with the aid of the secondary sources).

3.6 Bibliographic Control

As far back as the 19th century, the large amount of information resources available made it impossible for people to sift through and select only what was relevant to them. Some form of aid had to be devised when things were getting out of hand. Then publications intended to inform people of what information was available in various disciplines or that was published within national boundaries began to emerge. They came under such names as "bibliography", "indexes", "guide to the literature", and "directory". The term "bibliographic control" points to the fact that a specific service provides an information user with a comprehensive list of bibliographic records of the information sources that have come into existence within the period covered.

A bibliographic record for a book consists of the name of the author (or names of the authors), the title of the publication, the place of publication, the publisher and the year of publication. A record for a journal article consists of the name of the author (or names of the authors), the title of the article, the journal in which it was published, the issue of that journal and the pages taken up by the article.

The bibliographic record of a book will be written like:
O.O. Galu. *Democracy in Nigeria*. Lagos: Home Press, 2001.

The bibliographic record of a journal article will be written like:
C. P. Mogaji. “Life Cycle of Fleas.” *Top Journal of Zoology*, 34 (3), 1987, 56-89.

4.0 CONCLUSION

In this unit, you have learnt the concepts of information and document, the characteristics of documents, and the principles and history of documentation. You should now be able to explain howl research contributes to the growth of information and creation of knowledge.

5.0 SUMMARY

In this unit, you have learnt the basic concepts of information, document and documentation. These are fundamental to the whole business of information storage and retrieval. In the next unit you will be introduced to the organisation of information.

6.0 TUTOR-MARKED ASSIGNMENT

Write an essay on "Documentation and Growth of Information." Your write-up should be between six and eight pages of A4, typed with double spacing and 12 points Times Roman.

7.0 REFERENCE/FURTHER READING

Susan, A. (1972). *An Introduction to Computers in Information Science*. Metuchen, N.J.: Scarecrow.

UNIT 3 CLASSIFICATION

CONTENTS

1.0 INTRODUCTION

In the last unit, you learnt the basic concepts of information, document and documentation. Now you know what we store and retrieve in an information system. In this unit, you will be introduced to the organisation of information.

2.0 OBJECTIVES

At the end this unit, you should be able to:

- explain the meaning and purpose of classification
- describe several approaches to the creation of categories
- distinguish between natural and artificial classification
- describe the nature of hierarchical classification
- classify documents using a number of different classification schemes.

3.0 MAIN CONTENT

3.1 Need for Order

One of the capabilities of man is his ability to organise and see things in order. When you walk into a super-market you will see that the items there are displayed in sections. The items of the same are in the same section. Can you think of why that is so? You would have realised that it is much easier to locate the items. You would not be looking for shoes in a section with the label "household ware". When you go into the market, you do not go on roaming all over the place. You know exactly which area to go because you know what you need and you know which section of the market sells, things of that type. Perhaps the best example of organisation of materials you have noticed or may notice is in the library. We shall soon talk more about how such elaborate organisation is done.

First of all how do you arrange your books? It is very likely you do not arrange or organise, them in any particular order; and certainly you do not have any problem with that. You can pick out any book you want from the heap without any difficulty just because the books are not very many. Now imagine a room with piles and piles of books on the floor and almost reaching the ceiling. The whole room has been taken up by books. How easy would it be to locate a book from that room? Sure enough, nobody would like the unpleasant task of bringing books from there for people who come to ask for them.

In order to deal with this difficulty, various ways have been evolved to organise knowledge to classes representing subject fields. Each class is divided into subclasses dealing with recognisable segments of that field. The subclass is also divided into smaller units representing well-defined subject matter within the segments in a subclass. Information resources or documents are then arranged in classes, subclasses, and units within subclasses. Every document that is stored in the library is represented by a record in a collection of files, which is made available to those who come o seek for information.

3.2 What is Classification?

Classification is the act of grouping things together. Classification portrays the relationships between things, and between their classes. In fact, classification is a way of imposing order on creation. Since it is easier to think in terms of classes than individual things in creation, classification allows us to generalise. 1t would be difficult if not impossible to reason if human beings did not have the power of classifying and creating categories. What we know as knowledge is the outcome of grouping, dividing and registering thoughts, things and ideas in an unlimited number of ways.

For consistent classification, there must be a classification scheme. A classification scheme is simply an orderly arrangement of categories of classes, a class being any group of entities sharing the same characteristics. A characteristic is an attribute by which concepts may be separated into groups or further subdivided by subject. Thus, the purpose of classification is to bring together (or form classes of) entities that share common characteristics and to separate entities that do not share common characteristics.

3.3 Aristotle's Categories

According to Aristotle, all scientific knowledge consists of the arrangement of particulars under class concepts or universals, and in the combination of these concepts into a system. He saw the goal of science as being to define and explain the nature of a subject by its essential properties and by its differentiating properties, which set it apart from other groups. In other words, the goal of science is a complete classification of objects of knowledge into classes with the characteristic similarities within groups and differences between groups. Aristotle further stated that the definition of a term or a class concept must be a complete statement of:

(a) the essential attributes of the class
(b) the peculiar attributes of the class
(c) the next higher genus
(d) the properties which differentiate it from others
(e) accidents (that is, properties that are not part of the definition but common to the class and other classes).

In his attempt to classify universal knowledge, Aristotle created ten classes or categories of models of being. They are as follows:

1 Substance
2. Quantity
3. Quality
4. Relation
5. Place
6. Time
7. Situation or position
8. Possession or acquired character
9. Activity
10. Passivity.

3.4 Emmanuel Kant's Categories

The German Philosopher, Emmanuel Kant further elaborated Aristotle's ideas and defined four categories, which he regarded as the fundamental and universal forms of thinking objects and their relations. He pointed out that through the use of these categories, the mind builds up the material of sense perception into a systemised or orderly whole of intelligible experience. Kant's categories are as follows:

1. Categories of quantity
   Unity
   Plurality
   Totality
2. Categories of quantity
   Reality
   Negation
   Limitation
3. Categories of relation
   Inherence and subsistence, or substance
   Causality and dependence
   Community, or reciprocity of causal influence
4. Categories of modality
   Possibility - Impossibility
   Existence - Non-existence
   Necessity – Contingency.

3.5 Natural and Artificial Classification

Here we may wish to distinguish between natural and artificial classification. A natural classification exhibits the inherent properties of things being classified. It is based on the natural properties that occur regularly and cannot be separated from the things being classified. Such classification, which is said to conform to the order of nature, is also referred to as philosophical classification. Artificial classification, on the contrary, is based only on some accidental property of things. It is usually a case of grouping things for specific purposes on the basis of arbitrary selection of an accidental trait in the objects being classified.

There are different types of natural classification, which may be identified by the internal structure of the class classification. For instance, we may distinguish between hierarchical and referential classification. The latter is a pragmatic approach to classification using a single trait or property irrespective of other characteristics. The same thing may be classified differently depending on the property used. Here we are more interested in hierarchical classification.

3.6 Hierarchical Classification

The underlying assumption in hierarchical classification is that the process of subdivision must show the natural hierarchy of the subject, proceeding from classes of "greater extension and small intension to those of smaller extension and greater intension". Following this principle, Bliss in his classification put the general works first, following these with works on general subjects treated specially, then with works on special subjects treated generally; and lastly, with works on special subjects treated specially. Bliss designed the following format which shows graded specifications that are applicable to the systematic subdivision of most subjects, general or special.

General in scope

Bibliographical
Historical and critical
Historical
Method + scope, and relations of the subject to others
Critical
Biographical
Ancillary: statistics, illustration, etc. documents, reports, etc.
Miscellaneous
Periodicals and serials of societies etc
Collections, selections, readings, miscellanies, essays

General in scope and treatment

Elementary, introduction
Manuals, compends
Treatise, principles, comprehensive studies
Discourses

General in scope and special in treatment

Theoretical treatises
Aspects of general subject
Treatment for special purposes, interests, professions, etc
Technical
Experimental and laboratory

Special in scope and treatment

Special subjects
Special theories
Aspects in special interests
Special topics
Special methods, experiments, etc.
Statistical treatment
Pamphlets of special content, and other special materials.

The principles of hierarchical classification as proposed by Shera and Egan are summarised by Wynar as follows:

1. A hierarchical classification proceeds by assembling the groups of sciences of the principal fields of knowledge into main classes or divisions, which are dictated by the theory of knowledge accepted. Such classes have great extension and small intension.
2. The process is continued by the designation of differentiating qualities within each main class, and thus, subclasses or subdivisions are made.
3. Each subdivision in turn is divided by further differentiating qualities to produce still further subdivisions, and still others successively to make sections and subsections, until further subdivision is impossible or impracticable.
4. Every subdivision of a class is subordinate to the class heading. The sum of these subdivisions is the whole meaning of the class.

3.7 Dewey Decimal Classification

Some form of classification schemes have been used since ancient times for organising materials in libraries. They include chronological arrangement, arrangement by title, grouping by broad subject, as well as arrangement by author, order of accession, size, and so forth. The rapid growth of library collections and their use during the nineteenth century resulted in a pressing need for better methods of organising library collections. Dewey decimal classification, developed towards the end of the 19th century is one of the many well-known library classification schemes that follow the hierarchical approach described above. Dewey divided the whole of knowledge into 10 main classes as follows:

First summary: The 10 main classes

000 Generalities
100 Philosophy & related disciplines
200 Religion

300 The social sciences
400 Language
500 Pure sciences
600 Technology (applied sciences)
700 The Arts
800 Literature (belles-lettres)
900 General geography & history

The second summary is made of the 100 divisions as follows:
Second Summary: The 100 Divisions

000 Generalities

010 Bibliographies & catalogs
020 Library & information sciences
030 General encyclopedic works
040
050 General serial publications
060 General organisations & museology
070 Journalism, publishing, newspapers
080 General collections
090 Manuscripts & book rarities

100 Philosophy & related disciplines

110 Metaphysics
120 Knowledge, cause, purpose, man
130 Popular & parapsychology, occultism
140 Specific philosophical viewpoints
150 Psychology
160 Logic
170 Ethics (moral philosophy)
180 Ancient, medieval, Oriental
190 Modern Western philosophy

200 Religion

210 Natural religion
220 Bible
230 Christian doctrinal theology
240 Christian moral & devotional
250 Local church & religious orders
260 Social & ecclesiastical theology
270 History & geography of church
280 Christian denominations & sects
290 Other religions & comparative

300 The social sciences

310 Statistics
320 Political science
330 Economics
340 Law
350 Public administration
360 Social pathology & services
370 Education
380 Commerce
390 Customs & folklore

400 Language

410 Linguistics
420 English & Anglo-Saxon languages
430 Germanic languages German
440 Romance languages French
450 Italian, Romanian, Rhaeto-Romancc
460 Spanish & Portuguese languages
470 Italic languages Latin
480 Hellenic Classical Greek
490 Other languages

500 Pure sciences

510 Mathematics
520 Astronomy & allied sciences
530 Physics
540 Chemistry & allied sciences
550 Sciences of earth other worlds
560 Paleontology
570 Life sciences
580 Botanical sciences
590 Zoological science

600 Technology (applied sciences)

610 Medical sciences
620 Engineering & allied operations
630 Agriculture & related
640 Domestic arts & sciences
650 Managerial services
660 Chemical & relate technologies
670 Manufactures
680 Miscellaneous manufactures
690 Buildings

700 The arts

710 Civic & landscape art
720 Architecture
730 Plastic arts sculpture
740 Drawing, decorative &
750 Painting & paintings
760 Graphic arts Prints
770 Photography & photographs
780 Music
790 Recreational & performing arts

800 Literature (Belles-lettres)

810 American literature in English
820 English & Anglo-Saxon literatures
830 Literatures of Germanic languages
840 Literatures of Romance languages
850 Italian, Romanian, Rhaeto-Romancc
860 Spanish & Portuguese literatures
870 Italic languages, literatures, Latin
880 Hellenic languages literatures
890 Literatures of other languages
900 General geography & history
910. General geography, travel
920 General biography & genealogy
930 General history of ancient world
940 General history of Europe
950 General history of Asia
960 General history of Africa
970 General history of North America
980 General history of South America
990 General history of other areas

The third summary is the 1000 sections. The 1000 section for pure sciences is as follows:

Third Summary: The 1000 sections

500 Pure sciences
501 Philosophy & theory
502 Miscellany
503 Dictionaries & encyclopedias
504
505 Serial publications
506 Organisations

507 Study & teaching
508 Collections, travels, surveys
509 Historical & geographical treatment
510 Mathematics
511 Generalities
512 Algebra
5 13 Arithmetic
514 Topology
515 Analysis
516 Geometry
517
518
519 Probabilities & applied mathematics
520 Astronomy & allied sciences
521 Theoretical astronomy
522 Practical &spherical astronomy
523 Descriptive astronomy
524
525 Earth (Astronomical geography)
526 Mathematical geography
527 Celestial navigation
528 Ephemerides (Nautical almanacs)
529 Chronology (Time)
530 Physics
531 Mechanics
532 Mechanics of fluids
533 Mechanics of gases
534 Sound & related vibrations
535 Visible light & paraphotic
536 Heat
537 Electricity &electronics
538 Magnetism
539 Modern physics
540 Chemistry & allied sciences
541 Physical & theoretical chemistry
542 Laboratories, apparatus, equipment
543 General analysis
544 Qualitative analysis
545 Quantitative analysis
546 Inorganic chemistry
547 Organic chemistry
548 Crystallography
549 Mineralogy

SELF-ASSESSMENT EXERCISE

Using the Dewey Classification schedules above find the class code for the following documents:

i. Practical Lessons in Chemistry
ii. Report of the Agency for Agricultural Development
iii. The Main Political Events and Figures in Nigeria, 1900 – 2000

a) *Practical Lessons in Chemistry 542*
b) *Report of the Agency for Agricultural Development 630*
c) *The Main Political Events and Figures in Nigeria 320*

3.8 Library of Congress Classification

The Library of Congress classification was developed as a series of special classification schedules between 1899 and 1920. The scheme was not intended to be a philosophical classification, but a very practical tool, an enumerative classification whose schedules are based entirely on the subject grouping of the collection of books in the Library of Congress. The scheme uses both letters and numbers (mixed notation), unlike Dewey decimal classification, which uses only numbers. Single letters are assigned to the main divisions. The Library of Congress classification schedules are as follows:

A	General Works
B-BJ	Philosophy. Psychology BL-BX Religion
C	Auxiliary Sciences of History
D	History: General and Old World (Eastern Hemisphere)
E-F	History: America (Western Hemisphere)
G	Geography. Anthropology. Recreation
H	Social Sciences
J	Political Science
KD	Law of the United Kingdom and Ireland
KF	Law of the United States
L	Education
M	Music. Books on Music
N	Fine Arts
PA	General Philology and Linguistics. Classical Languages and Literatures
PA	Supplement Byzantine and Modern Greek Literature Medieval and Modern Latin Literature
PB-PH	Modern European languages
PG	Russian Literature
PJ-PM	Languages and Literatures of Asia, Africa.

	Oceania. American Indian languages. Artificial languages
PN, PR, PS, PZ	General Literature. English and American Literature. Fiction in English Juvenile Literature
PQ, Part 1	French Literature
PQ, Part 2	Italian, Spanish, and Portuguese Literatures
PT, Part 1	German Literature
PT, Part 2	Dutch and Scandinavian Literatures
Q	Science
R	Medicine
S	Agriculture
T	Technology
U	Military Science
V	Naval Science
Z	Bibliography. Library Science

3.9 Bibliographic Classification

Henry Evelyn Bliss published his major work, *A Bibliographic Classification*, in four volumes between 1940 and 1953. It is based on the 26 letters of the alphabet to cover all of knowledge, with numerals for indicating form of material. The main classes are shown below:

A.	Philosophy and General Science
B.	Physics
C.	Chemistry
D.	Astronomy, Geology, Geography
E.	Biology
G.	Geology
H.	Anthropology
1.	Psychology
J.	Education
K.	Social Sciences
L-O.	History, Social, Political and Economic, including Geography.
P.	Religion, Theology, Ethics
Q.	Applied Social Sciences
R.	Political Science
S.	Law
T.	Economics
U.	Arts: Useful Arts
V.	Fine Arts
W-Y.	Literature and Language
Z.	Bibliography, Bibliography, Libraries
1-9	Anterior numerical classes (for special collections)

3.10 Universal Decimal Classification

The Universal Decimal Classification (UDC) was first published in 1899. It owes a lot to Dewey decimal classification. It is in fact an expansion of the Dewey decimal classification. UDC was designed for subject indexing of all branches of knowledge, with decimal notations for specifying the level of classification. The main classes are as follows:

0. Generalities of knowledge
1. Philosophy, Metaphysics, Psychology
2. Religion Science
3. Social Science
4. Mathematics and National Sciences
5. Applied Science, Medicine, Technology
6. The Arts, Recreation, Entertainment, Sport
7. Literature, Bells-Letters, Phylology, Linguistics, Languages
8. Geography, Biography, History.

3.11 LISA Subject Headings

When detailed classification of a specific subject field is necessary, the classification schemes presented above are usually inadequate. It then becomes necessary to create an original scheme for the purpose. For instance, the Library and Information Science Abstracts (LISA), published by Bowker Saur uses its own classification scheme. The broad subject headings are as follows:

Broad subject headings

1.0 Librarianship and Information Science
2.0 Profession
3.0 Libraries and Resource centres
4.0 Library Use and Users
5.0 Materials
6.0 Organisations
7.0 Library Buildings
8.0 Library Technology
9.0 Technical Services
10 Information communication
11.0 Bibliographic control
12.0 Bibliographic records
13.0 Computerised information storage and retrieval
14.0 Communications and information technology
15.0 Reading
16.0 Media

17.0 Knowledge and learning
18.0 Records management
19.0 Other fringe subjects

Subdivisions

10.0 Information communication
10.1 Information work
10.11 Social sciences, business information work
10.12 Humanities information work
10.13 Science, technology, medicine information work
10.14 Information services
10.15 Reference work
11.0 Bibliographic control
11.1 Bibliography
11.11 Bibliographies
12.0 Bibliographic records
12.1 Periodicals control
12.11 Cataloguing and indexing
12.12 Cooperative cataloguing, bibliographic utilities
12.13 Cataloguing rules
12.14 Bibliographic description
12.15 Manual catalogues
12.16 Computerised catalogues
12.17 Online catalogues
12.18 CD-ROM catalogues
12.19 Indexing
12.2 Book indexing
12.21 Subject indexing
12.22 Searching
12.23 Index language and systems
12.24 Subject heading schemes
12.25 Thesauri
12.26 Classification
12.27 Classification schemes
12.28 Computer assisted indexing
13.0 Computerised information storage and retrieval
13.1 Economic and commercial aspects
13.11 Networks
13.12 Software
13.13 Automatic text analysis, automatic indexing, machine translation
13.14 Searching
13.15 Downloading
13.16 Databases in general
13.17 Non-bibliographic databases, databanks
13.18 Bibliographic databases
13.19 Image databases

13.2	Full text databases
13.21	Multimedia
13.22	Online systems
13.23	Online databases
13.24	Disc stored systems
13.25	CD-ROMs
13.26	CD-ROM databases
13.27	Other disc stored systems
13.28	Other storage systems
13.29	Videotex
14.0	Communications and information technology
14.1	Computer industry
14.11	Networks
14.12	Computer science
14.13	Computers
14.14	Software
14.15	Imaging technology
14.16	Online systems
14.17	Disc stored systems
14.18	Telecommunications and broadcasting technology
14.19	Computer applications
15.0	Reading
15.1	Literacy
16.0	Media 16.1 Copyright
16.11	Printing, publishing and bookselling
16.12	Printing
16.13	Printing history and analytical bibliography
16.14	Publishing and bookselling
16.15	Authorship
16.16	Publishing
16.17	Publications
16.18	Electronic publishing
16.19	Bookselling
16.20	Audiovisual materials
16.21	Broadcasting
17.0	Knowledge and learning
17.1	Research
17.11	Education
18.0	Records management
19.0	Other fringe subjects

SELF-ASSESSMENT EXERCISE

Using the subject headings above, find the class code for the following documents:

i. National policy on telecommunications

ii. A new technique for image processing
iii. The use of CD-ROM in distributing databases

a.	*National policy on telecommunications*	*14.18*
b.	*A new technique for image processing*	*14.15*
c.	*The use of CD-ROM in distributing databases*	*13.26*

4.0 CONCLUSION

In this unit, you were introduced to the organisation of information. Now, you should be able to explain the meaning and-purpose of classification, describe several approaches to the creation of categories, distinguish between natural and artificial classification, describe the nature of hierarchical classification, and classify documents using a number of different classification schemes.

5.0 SUMMARY

In this unit and the one before, you learnt what we store and retrieve in an information system, and you were introduced to the organisation of information. In next unit you will learn how to recognise at a greater level of detail the concepts treated in a document and translate them into appropriate tags for storage and retrieval purposes.

6.0 TUTOR-MARKED ASSIGNMENT

Write an essay on "Organisation of Information." In your write-up you are to consider the following points:

i. Need for organisation
ii. Approaches to creating categories
iii. Types of classification
iv. General nature of existing classification schemes
v. Need for improvising and improvement.

7.0 REFERENCES/FURTHER READING

Brian, B. (1979). *Theory of Library Classification.* London: Clive Bingley.

Bohdan S.W. (1976). *Introduction to Cataloguing and Classification.* (5th ed.). Littleton, Colorado: Libraries Unlimited.

Henry, E.B. *The Organisation of Knowledge in Libraries*. (1939). New York: Wilson.

Shera, J.H. & Egan, M.E. *The Classified Catalog: Basic Principles and Practice*. (1956). Chicago: American Library Association.

UNIT 4 SUBJECT INDEXING

CONTENTS

1.0 INTRODUCTION

In the last unit we addressed the need for organising information sources. We were actually, concerned with creating broad subject categories. In this unit, our attention will be on recognising at a greater level of detail, the concepts treated in a document and translating them into appropriate tags for storage and retrieval purposes.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain what is meant by subject indexing
- describe the process of content determination
- correctly identify concepts to use as index terms
- correctly assign index terms
- describe the general principles of automatic indexing.

3.0 MAIN CONTENT

3.1 What is Subject Indexing?

Information retrieval, especially from a computerised system, requires a great level of detail of recognition of issues dealt with in the document.

Subject indexing is about recognising the issues dealt with in a document and assigning them to appropriate classes. The purpose is to facilitate storage, location, and retrieval of information. Just as the organisation of document into subject categories makes it possible to avoid examining a large number of documents before one finds the one needed, so also subject indexing reduces the number of subject classes one has to search for some needed information.

Let us consider a simple overview of subject indexing. The intention is to identify the various key issues or subject matter or topics addressed in a document, assign each one to an appropriate class, to which a tag or label is given. In this way one document may be assigned to as many classes as the number of key issues identified in the document.

Subject indexing consists of three distinct operations, namely: content determination, concept selection, and assignment of index terms.

3.2 Content Determination

The first step in subject indexing is to determine what a document is all about. There is need to find out the key issues addressed for which users will require the document. How does the indexer determine this? He or she has to scan, first the title, then any of the following features that are available: the abstract, the contents, the foreword, and the preface, and finally the main body of the text. The issues or subject matter or topics dealt with in the document (otherwise called concepts) are identified in the process.

It requires somebody who understands the subject to undertake this analysis. A very important consideration is the community of users who are expected to come to request for information. The indexer really has to know the profile of the user community. In many cases, every information retrieval system is intended for a defined user community.

SELF-ASSESSMENT EXERCISE

Identify four concepts from the following title and abstract, which you consider the article has addressed.

Bishop, A.M. (2001). “The Effects of Fertilisers on Soils in the Rain Forest Zone of Nigeria.” *Journal of Agricultural Services,* 8(2), 2001, 23-33.

Abstract

The research reported here addressed the use of fertilisers by root-tuber farmers in the rain forest belt of Nigeria. The effects on soil types, especially on the consistency and chemistry, were determined using three different analytical techniques. A number of recommendations are made for more effective agricultural extension services and better soil management.

Here are the concepts that seem important:

- Use of fertilisers
- Effects of fertilisers
- Soil consistency
- Soil chemistry
- Agricultural extension services
- Soil management.

In recognising these concepts, we have at the back of our mind those who should see this article. We could add:

- Agronomy
- Root tubers
- Soil types
- Analytical techniques
- Rain forest zone
- Nigeria.

But these are more likely to be addressed in a superficial way. You may wonder why the term "agronomy" was selected. The information in this article actually belongs to the field of agronomy. Giving the tag "agronomy" will make it possible for anyone who needs general information in agronomy to get at this article.

3.3 Concept Selection

It is good to realise that concept selection is, to a large extent, subjective. The indexer has to use his or her judgment to decide what is important, taking into consideration the needs of the expected users. In the example above we chose the six concepts because, from all indications, these seem to be the main subject matter treated in the article. We are actually saying that anyone who needs information in each of these six topics should read this article. Generally, selection of concepts depends on:

- The background of the indexer
- The ability of the author to communicate his thoughts with clarity
- The profile of the user community
- The objectives of the system.

We did not have to stop at the six concepts. We could recognise as many as six more as listed above. However, it is not wise to just keep on adding more and more concepts. Every concept that we recognise and use as a means of access to the document has implication for storage space. Another problem is that some concepts that are treated marginally will, if included in the list of main concepts, create the problem of retrieving irrelevant materials. Now imagine that someone wants information on tertiary education in Nigeria. If the article above was assigned to the class Nigeria (that is if the concept "Nigeria" was recognised as one of the subject matters treated in the article), then a search for information on Nigeria will retrieve this article along with all other articles for which the concept "Nigeria" was also recognised. Do you think the article would be useful to that person? Of course not.

SELF-ASSESSMENT EXERCISE

What do you think would be wrong with choosing more than the six concepts listed and including possibly up to six more?

3.4 Assignment of Index Terms

The issue here is the choice of appropriate terms to use for representing the concepts that have been selected. There are different approaches. One approach is to use the terms as they occur in the, document. This has the appeal of being natural and easy. It also has its problems, which will be pointed out in due course. Another approach is to translate the concepts into equivalent predetermined terms, descriptors,, on subject headings, classification numbers or other codes, that are used in a given information retrieval system. Whatever terms, descriptors, subject headings, classification number and other codes, (whether they occur naturally in the document or they were derived from translation process), that are used as tags for the document become the surrogates (representatives) of the actual documents.

SELF-ASSESSMENT EXERCISE

Once again, go through the title and abstract of the article by A. M. Bishop. Below are 11 concepts chosen to represent the article; identify which ones are natural and which ones have been translated.

A.M. Bishop. “The Effects of Fertilisers on Soils in the Rain Forest Zone of Nigeria.” *Journal of Agricultural Services,* 8(2), 2001, 23-33.

Abstract

The research reported here addressed the use of fertilisers by root-tuber farmers in the rain forest belt of Nigeria. The effects on soil types, especially on the consistency and chemistry, were determined using three different analytical techniques. A number of recommendations are made for more effective agricultural extension services and better soil management.

Concepts that have been chosen:

- Use of fertilisers
- Effects of fertilisers
- Soil consistency
- Soil chemistry
- Agricultural extension services
- Soil management
- Agronomy
- Root tubers
- Soil types
- Analytical techniques
- Rain forest zone

See the table below for the answer

Natural	**Translated**
Use of fertilisers Effects of fertiliser Agricultural extension services Analytical techniques Rain forest zone	Soil consistency Soil chemistry Soil management Agronomy Root tubers Soil types

3.5 Automatic Indexing

It has become possible to reduce the intellectual effort by human beings in indexing with the use of the computer. Now computers are being used not only to index documents but also to create links between documents for the benefit of the user of an information retrieval system.

There are several approaches to automatic indexing. They include word extraction by frequency, word position, word type as well as assignment indexing.

3.5.1 Word Extraction by Statistical Criteria

In this approach, words are extracted from the text on the basis of frequency of occurrence. This technique was pioneered by Luhn in 1957 and Baxendale in 1958, both of IBM. The computer counts the words and phrases that occur in the document and select the words that occur most frequently as the index terms. A stop list is used to exclude common, non-substantive words; that is all those words that are not useful for representing the subject matter in a document. This method has over the years proved to be reliable and effective.

Stop words

In an information retrieval system for agriculture, we would expect the following words to be among the stop words:

all
be
become
have
in
if
like
no

for
to
more
now
here
go
take
again

bring
often
always
many
out
near
produce
make

While this method is based on the absolute frequency of occurrence, another technique uses relative frequency of occurrence. In the letter, a word is extracted if it occurs more frequently than an expected value in the context of that subject.

3.5.2 Word Extraction by Non-Statistical Criteria

The use of non-statistical criteria is another approach, which may be combined with or replace the extraction by the statistical criteria. The extraction program evaluates the significance of words in the document on the basis of their position, the type of word, as well as the emphasis on a word as indicated by print features such as font size, boldface, italics and so forth.

3.5.3 Assignment Indexing

In this approach the indexing program assigns to a document one or more terms from a comprehensive list of terms called the vocabulary of the retrieval system. The first step is to identify the most appropriate terms from the document using any of the extraction methods described above. The selected terms are matched against the list of terms in the controlled vocabulary of the retrieval system. The terms in the vocabulary that best match the extracted words (that is, the best equivalents) are then used as the index terms.

SELF-ASSESSMENT EXERCISE

Read over the section and make sure that you can explain the three approaches to automatic indexing, namely:

- Word extraction by statistical criteria
- Word extraction by non-statistical criteria
- Assignment indexing.

3.6 Citation Indexing

Citation indexing is not part of subject indexing, but it will be described briefly because of its, importance in information storage and retrieval. The citation index is a list of references, each of which is accompanied with a list of documents that have cited it. It is assumed that bibliographic citations provide some indication of a relationship between documents especially in terms of subject content. A good example of citation index is the one being produced by the Institute of Scientific Information, Philadelphia, in the United States.

4.0 CONCLUSION

In this unit you have learnt both the principles and procedures of subject indexing. By now you should be able to explain what is meant by subject indexing, describe the process of content determination, correctly identify concepts to use as index terms, and describe the general principles of automatic indexing.

5.0 SUMMARY

In this unit, you have learnt to recognise the main concepts treated in a document and translate them into appropriate tags or index terms for storage and retrieval purposes. In the next unit, you will learn something about indexing languages and their characteristics that determine how such concepts are translated into index terms.

6.0 TUTOR-MARKED ASSIGNMENT

Describe the general principles and procedures in subject indexing. Your write-up should be between six and eight pages of A4 typed with double spacing in 12 points of Times Roman.

7.0 REFERENCE/FURTHER READING

Lancaster, F. W. (1979). *Information Retrieval Systems: Characteristics, Testing and Evaluation*. New York: Willey.

UNIT 5 INDEX LANGUAGE

CONTENTS

1.0 INTRODUCTION

In the last unit, you learnt the basic principles and procedures of subject indexing. Among the things you learnt is the selection of the main concepts treated in a document and their translation into appropriate tags or index terms. In this unit, you will be introduced to indexing languages and their characteristics that determine how such concepts are translated into index terms.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain what is meant by indexing language
- distinguish between natural and controlled language
- explain the purpose of controlling vocabulary
- explain the use of a thesaurus
- describe pre-coordinate system
- describe post-coordinate system.

3.0 MAIN CONTENT

3.1 What is an Indexing Language?

Language of a system is the total number of index tags available for use in the subject description of documents. There are two basic types of indexing languages: the natural language and the controlled vocabulary. The three-step indexing activity described in the previous unit presumes a controlled vocabulary. The third step was the translation of terms or concepts selected from the document into the most appropriate equivalents that are available in the vocabulary of the system. Where there is no such standardised vocabulary, the third step is irrelevant.

A natural language is one in which there is no restriction on terms to be used as long as they, are substantive terms used in the document. Natural language indexing is also called free-text indexing. The use of terms in the document gives high specificity, which in turn leads to the retrieval of very relevant information. There is a greater level of exhaustivity and this tends to retrieve as much information as possible. When new words are coined by authors, they come immediately into the system and the system is usually quite up-to-date.

The need to search for information across a number of databases in one search operation has made natural language indexing far more promising than controlled vocabulary. Natural language, allows easy retrieval from one database to another without any problem of incompatibility of terms.

The appeal of natural language indexing has increased with the success of automatic indexing. It is very easy to use computer programs to identify substantive terms that qualify for use as index tags and pick them out. Indexing cost is now considerably low with automatic indexing. The advantages of a natural language are:

- simple and easy indexing procedure
- requires no subject expert
- freedom of choice of terms
- high specificity ensures retrieval of relevant documents
- exhaustivity ensures the retrieval of a good number of documents
- allows transparent searching of multiple databases in one search operation
- high efficiency and effectiveness with automatic processing.

SELF-ASSESSMENT EXERCISE

Now study the two titles below and try to identify the appropriate natural language terms.

i. Post-Harvest Storage of Guinea Corn
ii. How to Store Sorghum after Harvest

3.2 Controlled Vocabulary

Natural language indexing has its own limitations. One obvious limitation is the scattering of information under different synonyms, different forms of the same word, and so forth. Another limitation is its inability to supply index terms that are implied in the context but not expressly used in the document.

SELF-ASSESSMENT EXERCISE

What is the problem with the natural terms that we identified in the previous self-assessment exercise?

It would be obvious to anyone that the two documents are related and so should go together. According to natural language indexing the first one would be represented by: post harvest, storage and guinea corn while the second would be tagged with: store, sorghum and harvest. What has happened is that the two documents have been scattered instead of being brought together in the interest of the user. Even though the two documents are likely to be useful to a researcher working on cereals, natural indexing would not generate the term "cereals” since the term is not used in the document.

Controlled vocabulary has the advantage of:

- imposing order in the use of terms
- imposing economy by restricting the size of vocabulary and reducing the number of unique terms to be entered into the system
- providing a means of showing the relationships between terms
- standardising descriptors for both the indexer and the searcher
- ensuring (or increasing the probability) that the user who searchers the system will retrieve all the documents in the system that are relevant to his request.

The control of vocabulary is basically aimed at controlling:

- the size of the vocabulary
- the specificity of terms
- the exhaustivity of indexing
- the relationship between terms
- the relationship among words in compound terms (citation order)
- synonyms by using a single term to represent all its synonyms
- the form of word to be used.

3.3 Size of Vocabulary

The size of the vocabulary, that is, the number of index terms available for use in a specific retrieval system, is a very important factor in its retrieval performance. It has an economic cost in terms of storage space. It is also an important determinant of the quality of searching output, as it is related to other indexing considerations, especially specificity. This matter will be clearer when we deal with specificity.

3.4 Specificity of Terms

Specificity refers to the extent to which the indexer can select narrow terms (species) rather than the, broader more inclusive term (genus).

SELF-ASSESSMENT EXERCISE

Read the three titles below and think of a generic (broader) term for the underlined terms:

1. Cultivation of Rice in Swamps
2. A New Variant of Millet
3. Post-Harvest Storage of Maize

The three terms, namely rice, millet, and maize may be represented by the broader term "cereals". If we do so, then there is a loss of specificity. We cannot search for these documents with the specific terms, rice, millet or maize. The only means of retrieving them is by using the generic term "cereals". The size of the vocabulary is reduced also because a generic term is used to represent all its specific terms.

SELF-ASSESSMENT EXERCISE

What is the generic form of the underlined term in the following titles?

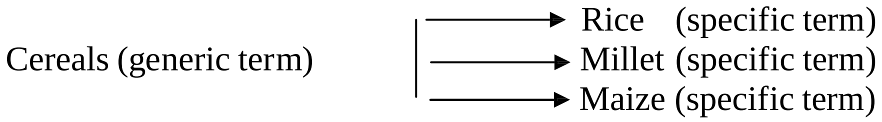
i. Learning the Art of Systems Programming
ii. Introduction to Programming in Pascal

3.5 Exhaustivity

Exhaustivity is the degree to which the indexer selects the different concepts dealt with in the document. The highest level of exhaustivity is achieved when all the concepts that could be identified in a document are selected. This level of exhaustivity is not always desirable. There is the consideration of optimum number of terms to select and the implications of that for the performance of the retrieval system. As a rule, it is only the concepts that are given sufficient coverage and warrant drawing user's attention to such information that should be selected.

Relationship between terms

Earlier we described "one kind of relationship among terms, namely the genus-species relationship. We took as our example the relationship between cereals as a genetic term and rice, millet and maize as specific terms.



Cereals (generic term)

Rice (specific term)

Millet (specific term)

Maize (specific term)

The relationship is also hierarchical, because the generic term which is a higher and more inclusive term can be broken into a number of lower-level specific terms.

Other hierarchical relationships include:

- a thing and its parts
- a thing and its processes
- a thing and its properties
- a thing and the operations performed on it.

SELF-ASSESSMENT EXERCISE

An example of each of these relationships is presented below. Try to identify the elements of the relationships:

- computer accessories
- data collection
- melting point of iron
- treatment of malaria

We can illustrate these relationships in a table.

Thing	Lower level	Relationship
Computer	Computer accessories	Parts
Data	Data collection	Process
Iron	Melting Point of iron	Property
Malaria	Treatment of Malaria	Operation

Citation order

Here the concern is how, to arrange the constituent words in a compound term in translating a concept - into an index term. Consider a document entitled "A Report on University Education in Nigeria". One of the concepts to be selected is "university education". In a system in which the order of citation, (that is, which word comes first) is important, we would have to decide which of the following versions to adopt:

- university education
- universities, education
- universities – education
- education, universities
- education - universities

Synonyms

Synonyms are words that have the same meaning. Words that are close in meaning (near synonyms) are also treated as synonyms. The problem with synonyms and near-synonyms is that they scatter documents in different places. That is not in the interest of the users who come to search for information. A decision has to be made which of the words to use to represent all the others as an index term. If a user comes to search with any of the non-index terms, the system should tell him or her the appropriate term, because a see reference is made from each of the synonyms that is not used as index term to the preferred term.

Form of words

Variant forms of words pose the same problems as synonyms. Now inspect the list of words below:

Computer	computers, computing
Process	Processing, processed
Graphic	Graphics
Finished	Finishing
Carbonated	Carbon
Module	Modular

Each word can be used as an index term but a control would mean choosing which one to represent all its variant forms.

Thesaurus

The thesaurus is a vocabulary of controlled indexing language, formally organised so that **a priori** relationships between concepts are made explicit. The index language of a system with a controlled vocabulary is usually set out in lists, which may be in the form of subject headings, thesaurus, or some classification schedule. The use of thesauri and their construction have been subjects of much interest. They were meant to be a tool to aid the indexer in the choice of index terms for consistency and to assist the searcher in using the same terms as the indexer for maximum retrieval results.

Pre-coordinate system

A pre-coordinate system is one in which terms are combined or arranged (coordinated) at the time of indexing to form the index terms. Once such terms have been made up, they are to be used by both indexer and searcher in the form in which they have been coordinated. Two important considerations, in pre-coordinate indexing are:

i) a decision on citation order as described above

ii) anticipation of the approaches the users will adopt in their search for information

Here is the title of a document for consideration.

“A Federal Government Report on University Education in Nigeria".

In a pre-coordinate system, the indexer will have to think of the possible search strategies of users and provide access to the document accordingly. The index tags that he will choose are the access points to the document. He will probably provide the following index terms.

- Education, Federal Government Report
- Education, Nigeria
- Education, tertiary
- Education, university
- Nigeria, Education
- Nigeria, universities
- Universities, Federal Government Report
- University education

The searcher has to use any of these terms exactly as it has been coordinated otherwise he will not hit this document.

SELF-ASSESSMENT EXERCISE

Can you think of the possible disadvantages of a pre-coordinate system? Just try.

There are two main limitations of a pre-coordinate system.

i. It imposes considerable rigidity in the use of terms both at the indexing and searching stages.
ii. It involves multiplication of index terms and so increases the size of the vocabulary and operational cost.

This problem may be reduced by reducing the number of terms and referring users from other possible forms to the selected forms. Pre-coordinate indexing is now really a feature of manual systems.

Post-coordinate system

In a post coordinate system, the combination of terms is not done at all during indexing. This leaves the searcher full freedom to choose and combine terms according to the information he wants. In the document "Federal Government Report on University Education in Nigeria", the most important terms that need to be stored are:

- Education
- Nigeria
- Report
- University

The user is free to combine any number of these terms to conduct his search.

SELF-ASSESSMENT EXERCISE

A list of topics is presented below. Reduce each topic into appropriate keywords:

- for a pre-coordinate system
- for a post-coordinate system
- national policy on science and technology
- progress in development of small-scale industries
- encouragement of indigenous inventions and innovations
- establishment of Raw Materials Research and Development Council
- primary health care and community development.

4.0 CONCLUSION

In this unit, you have learnt two main approaches to indexing namely, the use of natural language and the use of controlled vocabulary. You have learnt the characteristics of these languages that determine the choice of index terms and consequent retrieval performance. Now, you should be able to explain what is meant by indexing language, distinguish between natural and controlled language, explain the purpose of controlled vocabulary and thesaurus, and describe pre-coordinate and post-coordinate systems.

5.0 SUMMARY

You have now been introduced to indexing languages and their characteristics that determine how the concepts addressed in a document are translated into index terms. In the next unit, you will learn and appreciate the role of computers in information storage and retrieval.

6.0 TUTOR-MARKED ASSIGNMENT

Write a short essay on "The Importance of Indexing Language in Information Storage and Retrieval." Your write-up should be between four and six pages of A4, typed double-spaced with 12 points Times Roman.

7.0 REFERENCE/FURTHER READING

Susan, A (1972). *An Introduction to Computers in Information Science*. Metuchen, N.J.: Scarecrow.

MODULE 2

UNIT 1 COMPUTERS IN INFORMATION STORAGE AND RETRIEVAL

CONTENTS

1.0 INTRODUCTION

In the last unit you were introduced to indexing languages and their characteristics that determine how the concepts addressed in a document are translated into index terms. In this unit, you will learn and appreciate the role of computers in information storage and retrieval. You will also learn the basic architecture of a computer system.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain the strengths of the computer
- explain the history of the development of computers

- describe the organisation of a computer system
- describe the input and output devices
- distinguish between hardware and software
- explain the general principles of computer communication.

3.0 MAIN CONTENT

3.1 The Strengths of the Computer

The role of computers in information processing has become so important that organisations are investing substantial proportions of their income in acquiring and maintaining them for better performance in information processing. Computers are just fantastic when it comes to speed of processing data and information. In less than a minute, a computer can do a data processing job that will take five workers ten weeks to complete. The various computations and transformation of data actually take only a small fraction of a second. For the rest of the time, the computer is idle waiting to receive data.

Computers are very accurate in their operations. You may have heard people talking of computer error. There is nothing like that. Whatever error you observe in a computer operation is the error introduced by somebody either in the data that were fed into the computer or the instructions given to it to work with. When you work manually with a long list of figures, you would normally, repeat the computation several times to be sure that you have got the right result. It is not unusual to obtain different results as such a computation is repeated. In that case, it is usual to accept, as the correct result, the figure that turns up more frequently than others. With a computer you are sure that the result you get is reliable so long as the data and instructions supplied to the computer are in order.

The computer does not suffer from fatigue and distraction, and so it can work with sustained ability. This is not the case with human beings working manually on similar tasks. A very important feature of the computer is its capacity for storing data and information. If you visit some university libraries you will find that all the bibliographic records stored in cards in cabinets that previously filled a large hall have been transferred to a computer. Now, people are envisaging the time when all the information resources in the world will be in computers and will be accessible to everyone in the world. With the large potential of storage space in the computer, more and more information in the existing physical formats are being converted to electronic format for computer storage in order to save space and for easier dissemination through computer networks.

It is not interesting doing a repetitive cycle of data processing job. It gets boring. Now, such jobs are better given to, a computer to do. Then people are free to work on more creative jobs and develop their intellectual capabilities. When the first general-purpose digital computer was developed in 1944 in the United States of America, it was used only as an aid to computation.

SELF-ASSESSMENT EXERCISE

What are the capabilities that make computers so important in the processing of data and information?

3.2 Historical Background

After centuries of attempts to invent machines to aid computation, Charles Babbage began in 1830 the construction of his Difference Engine, which he expected automatically to compute and print mathematical tables. He abandoned the project in 1834 to start on his new idea of an Analytical Engine. The organisation of the Analytical Engine that he had in mind closely resembles that of modern digital computers. Unfortunately, he never completed the project.

The invention of the computer in the early years of the twentieth century was a direct response to the need for a faster and more efficient means of dealing with the enormous task of counting and analysing the United States population every ten years. In March 1884, a young scientist, Herman Hollerith obtained the first patent for a data processing machine. He invented a series of machines, which could accept data and record them in cards. The breakthrough consisted in entering data only once, after which they could be used again and again to generate reports according to different criteria. By 1889, his machines had been well proved and were chosen by the United States government for the 1890 census. With Hollerith's machines, the time to complete the results of the 1890 U.S. census was reduced to two and half years as compared to the seven and half years it took to compile the 1880 census.

In 1896, Hollerith started the Tabulating Machine Company. After a series of mergers, it became the Computing-Tabulating-Recording Company. In 1924, it was renamed International Business Machine Corporation (IBM). Hollerith's successor at the Census Bureau, James Powers, improved on Hollerith's machines. On leaving the Census Bureau, he also formed a company, which through mergers became Sperry Rand, the manufacturer of UNIVAC computers.

With substantial support from IBM, Howard Aiken of Harvard University developed in 1944 the first successful general-purpose digital computer called the Harvard Mark I. That was the first electro-mechanical computer. Atanasofs ABC (Atanasof-Berry Computer) was the first to use vacuum tube, thus becoming the first truly electronic digital computer. It was also the first to use the binary system for representing numbers. The ENIAC (Electronic Numerical Integrator and Computer) developed by John W. Mauchly and J. Presper Eckert, Jr. with financial assistance from the U.S. Army, was completed in 1945. Joined by John von Neumann, Mauchly and Eckert began working on a new version of ENIAC called EDVAC (Electronic Discrete Variable Automatic Computer) in 1946. About the same time, Maurice Wilkes began to develop the EDSAC (Electronic Delay Storage Automatic Calculator) at the University of Cambridge in the United Kingdom. Like the EDVAC, the EDSAC incorporated the concept of stored program. The EDSAC became operational in 1949 ahead of EDVAC. In 1951, the UNIVAC I from Remington Rand made its debut as the first commercial computer. Thereafter, IBM achieved prominence in the production of computers.

3.3 Generations of Computers

The period between 1944 and 1959 was that of first generation computers. The machines were of huge size. They were noisy and they generated much heat. The key electronic component was the electronic tube or valve. The most important memory material was magnetic core. The computers could process only a few thousand instructions per second and store between 10,000 and 20,000 characters.

The second generation came up between 1959 and 1964. The machines were based on transistors rather than on valves, and so they were faster, and more reliable and smaller in size. Improved techniques for using the machines were made possible through the development of operating systems, time-sharing technique, and introduction of high level programming languages. Both the first and the second generations featured machines of large size called mainframes. During this period, IBM rose to prominence.

The third generation (1964 - 1970) took off with the advent of printed and integrated circuits and production of the relatively smaller computer called mini-computers. They were faster, more compact, more reliable and cheaper than the giant-size computers of the previous generation. So, they had an edge over mainframes and they quickly dominated the market. That was the period of IBM's System/360 and Digital Equipment Corporation's (DEC) PDP-8.

The fourth generation is that of microcomputers. The history of microcomputers goes back to the 1970s. Further reduction in the size of computers became possible through large-scale integration (LSI) in which the circuitry of major components of the machines were telescoped into a single silicon chip that is smaller than half an inch square. Progress in microelectronics continued with growing ability to miniaturise circuits and to pack them at greater densities. Along with that ability was the growing storage capacity and speed of processing. Magnetic core memories were replaced by faster and cheaper metal oxide semiconductor (MOS) memories. Functionality was improved through progress in software engineering.

The popularisation of computer applications was aided by the fact that by the 1980s microcomputers became affordable to individuals. Furthermore, wide spectrum of applications was developed for all kinds of needs in real life situations. Emphasis on user-friendliness in software development produced applications which could be used with ease by people who were not computer specialists.

The fifth generation may be said to have started from 1990. The main developments that have been recorded since then include the supercomputers, further reduction of the size of computers to produce laptops and notebooks, software that incorporate a high level of intelligence and provide a high level of user-friendliness with graphical user interface. Remarkable improvements in the capabilities of microcomputers have been made, including clock speed that is well over 800 mega hertz, main memory that can store over 250 million characters, and hard disk capacity of over 40 billion characters.

3.4 Computer System

A computer system consists essentially of a central processing unit (CPU), a main memory, and various input and output devices. The central processing unit is to the computer what the brain is to a human being. It is responsible for coordinating the functions of all other components of the computer system. It carries out the arithmetic and logical functions of the computer. The CPU is made up of two parts: the control unit and the arithmetic and logical unit (ALU). The function of the control unit is to get a single instruction and decode it. After decoding the instruction, the control unit gives way to the ALU to execute it. Once the ALU completes the execution of the instruction, the control unit again takes over.

The main memory, otherwise called primary memory, is the area of the computer in which data and instructions are stored while the computer is working. A small part of the main memory is reserved for storage of instructions that the computer needs for organising itself. When you switch on a computer, the screen lights up and begins to display a series of information and messages, some of which show what the computer is doing. At this time, the computer is said to be "booting". It is organising itself to start work. It checks the main memory and the devices that it expects to have been connected. The instructions that it is using at this time come from what is stored in the part of the main memory called the ROM (read only memory). The instructions in ROM are kept alive with a backup battery.

The rest of the main memory is called RAM (random access memory). It is in the RAM that data and instructions are held while the computer is executing a job. The computer stores and manipulates data and instructions in two states called 0 bit and 1 bit. Eight bits form a byte, and a byte of different combinations of 0 and 1 bits is used to represent a character. All the characters, including the numerals 0 to 9, the, alphabets a to z and A to Z, as well as special characters are represented by combinations of eight bits. All that the computer does is done by the passage of a stream of pulses of electric current to create these bits.

SELF-ASSESSMENT EXERCISE

Switch on a computer. As it boots, note the messages.

3.4.1 Input/Output Devices

In order to get a computer to work for you, you need to give it some instruction. There are several ways to do so. You can use the keyboard or mouse or any other device. The keyboard is particularly suitable for giving a command. When you strike a key, a set of eight bits that is unique to that character is sent to the computer and it is stored in RAM. If you are typing a command, the computer just keeps on recording the characters until you strike the "enter" key, which signals the completion of the command. Then the control unit of the CPU fetches the instruction from the memory, decodes it, and prompts the ALU as well as other appropriate parts of the computer to act.

The mouse is meant for picking tasks from a menu. A menu in a computer application is a list of tasks that a person can choose from. Each task is also represented by a small picture frame called an icon. You can click on an item on the list or on an icon that represents that task. That tells the computer what you want it to do and it will begin to do it. One thing you will find quite interesting is the possibility of working on several screens and on different tasks at the same time by switching from one screen to another.

If you take a close look at the keyboard, you will see that the rows of keys marked F1 to F12. They are called function keys. In every application, certain tasks are tied up with these keys. When you strike one of them it will invoke the task tied to it and the computer will execute it.

You can also scan a document whether of text or graphics into the computer with the aid of a scanner. You can create images in the computer by using various devices. A camera may also be used to input an image into a computer.

The disc drives can be regarded as input devices when the computer is made to take, instructions and data from them. There are three types of drives: the hard disk drive, the floppy disk drive, and the CD-ROM drive. The hard disk drive normally stores the computer programs of the applications that are available in a computer system. It also provides a large storage space for the users' data and user-created programs. The floppy disk and CD-ROM provide facilities for a user to bring inputs to the computer in whatever formats, including text, graphics and sound, as well as to take outputs away. The difference between the two is that the floppy disk is a magnetic medium while the CD-ROM is an optical medium. We shall examine these media in greater details.

The output devices are those in which the results of computer operations are displayed or stored. When you strike a key on the keyboard you will see that character registered, or else you will see some action initiated. The screen displays whatever the computer does. The equipment that monitors what is going on in the computer and provides a screen to view it is appropriately called the monitor. The output of a, computer operation may be sent to a printer to be printed on paper, or sent to be stored on floppy disk, CD-ROM, plotted to produce a graph, or to one of several other media and devices.

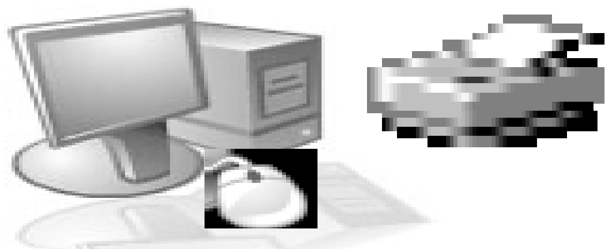


Fig. 1: A Computer System with Monitor, Keyboard, Mouse, Floppy Disk, Hard Disk and Printer

3.4.2 Hardware and Software

Hardware is obviously the physical equipment that constitutes part of a computer system. They include the components of a computer, the monitor, keyboard, printer, scanner, mouse, loudspeakers and other devices that may be connected to a computer.

Software is a collection of the computer programs that a computer requires to operate. A program is itself a set of instructions. Every task that the computer has to execute is supported by an, elaborate set of instructions. A computer cannot operate unless it has very clear, step by step, instructions on what it is to do. This means that a computer cannot act on its own initiative and cannot do what no human being knows how to do. Usually software is made up of a collection of programs for the specific tasks in an application.

Software is of two categories: the system software and applications software. Systems software is the one that a computer uses to manage itself and its peripheral devices (the input, output, and other devices connected to it). Systems software is to serve as intermediary between the physical devices of the computer, and the programs you choose to use to do your work. The systems software that you are more likely lo come across includes MS-DOS (not as popular as it used to be); Windows, Windows NT, UNIX, and some network software like Novell Netware.

Applications software are the collection of programs that guide the computer on the, execution of the job you want it to do for you. If you want to type a letter, you should select one of the word processing software such as Microsoft Word or WordPerfect. If you are going to produce elaborate tables of data with some columns whose values will be automatically computed according to your wish, you should select one of the available spreadsheet software. If you are using Windows, you, will most certainly have the Microsoft Office Professional and so the Microsoft Excel software for doing your spreadsheet.

SELF-ASSESSMENT EXERCISE

Find a place where you can have access to a computer and ask to be introduced to Windows.

3.5 Computer Networks

Computers are usually linked together in a network for the purpose of data and information communication and sharing. When the computers that are linked together are located in one building or in contiguous buildings, the network is referred to as a local area network (LAN). Special cabling has to be done to link the computers together. Communication of signals through the network is done in the digital mode, the same mode in which a computer operates.

There are frequent occasions when a computer or a local network has to be linked to a distant computer or network. Then it is cheaper to .use the telecommunication facilities already in place to effect the link. Most national telecommunications facilities transmit signals in analogue form. Analogue signals are not compatible with the digital signals of a computer. Therefore, a special device called modem (modulator/demodulator) is necessary between the computer and the telecommunication facility to superimpose the digital signals of the computer on the analogue signals of the telecommunication line (modulation) so that it can be transmitted in an analogue form. When the modulated carrier waves (that is the telecommunication signals carrying the computer signals) arrive at their destination, another modem recovers the computer signals from the carrier waves (demodulation) and inputs them into the destination computer or network. Through this link it is possible to store information in any computer anywhere in the world and make it accessible to individuals and organisations throughout the world. In fact, the business of managing data and information in enterprises now depends heavily on access to relevant computer networks and communication of data and information through networks.

3.6 Networks for Information Retrieval

Experiments with the use of computers to perform library operations began in the early 1960s id several North American and British libraries. In 1961 H.P. Luhn of IBM developed programs for producing indexes. In the same year catalogue cards were produced at the Douglas Aircraft Corporation with a computer. In the mid 1960s four libraries in London used the computer to produce a union catalogue of their holdings.

The 1970s ushered in greater success both in the development of integrated library applications and exchange of data. The MARC format and the ISBN were instrumental to the growth of cooperative services. The successes in developing computer-based retrieval systems in the early 1960s could be credited mainly to organisations in the United States of America. These include the Armed Services Technical Information Agency (later the Defence Documentation Centre) in the period 1959 - 1963, the National Aeronautics and Space Administration in 1962, and the National Library of Medicine, whose MEDLARS service was launched in 1963. (MEDLARS means Medical Literature Analysis and Retrieval System).

During the 1970s efforts towards cooperative services yielded the Birmingham Libraries Cooperative Mechanisation Project (BLCMP), the South-Western Academic Libraries Cooperative Automation Project (SWALCAP), the Ohio College Library Centre (OCLC, later renamed

Online Computer Library Centre), the Research Libraries Information Network (RLIN), the University of Toronto Library Automation System (UTLAS), and Washington Library Network (WLN).

The early information retrieval systems were exclusively for producing bibliographic records for distribution in printed form. Later many of the databases holding these records became available online (that is, through computers elsewhere logged into the host computer in which the records were stored). Now, most of these systems contain varying proportions of full-text articles in addition to the bibliographic information and abstracts. Efforts are being made to provide links between documents so that one document can lead a user to other related documents.

4.0 CONCLUSION

Conclusion In this unit you learned something about the history of the invention of the computer, the basic organisation of a computer system, and the role of computers and computer networks in information retrieval. Now that you have complete this unit, you should be able to explain the strengths of the computer and the history of the development of computers, describe the organisation of a computer system and the input and output devices that constitute the peripherals, distinguish between hardware and software, and explain the general principles of computer communication.

5.0 SUMMARY

In this unit, you have learnt among other things the basic architecture of a computer system and you are now in a position to appreciate the role of computers in information storage and retrieval. In the next unit, you will learn more about the media on which information is stored.

60 TUTOR-MARKED ASSIGNMENT

Write an essay on the contribution of computers to the development of information retrieval networks. Your write-up should be between four and six pages of A4, typed double-spaced with 12 points Times Roman.

7.0 REFERENCES/FURTHER READING

Cyril, H. P. *et al*. (1982). *Information Systems Design*. Sydney: Pretence-Hall.

French, C. S. (1996). *Computer Science*. (5th ed.). London: Letts Educational.

UNIT 2 STORAGE MEDIA

CONTENTS

1.0 INTRODUCTION

In the last unit, you learnt among other things the basic architecture of a computer system and you could appreciate the role of computers in information storage and retrieval. In this unit, you will learn more about the media on which information is stored.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain how the computer represents data
- describe the technical features of a number of storage devices.

3.0 MAIN CONTENT

3.1 Data Representation

In order to get the computer to work for you, you need to present to it both data and instructions in a form it can recognise and accept. The computer accepts and records data and instructions in 0 and 1 bits. This is called binary representation. A combination of eight of these bits makes one byte, which is the unit of representation of a character. Whatever data, information and instructions will be given to the computer must be converted to the binary form. You do not need to worry about how to effect, this conversion. There is always an interface between the input device and the computer to do the conversion.

Almost any device that can hold two states, such as current and no current, open switch and: closed switch, and so forth, could be used to store these bits; and so, is a potential material for manufacturing a computer storage device. We shall now describe some of the storage devices.

3.2 Magnetic Tape

A reel of magnetic tape is 10.5 inches in diameter. Magnetic tape came as a replacement for both punched cards and paper tape. In it data are represented through the arrangement of magnetised spots along the length of the tape, in rows referred to as tracks. Each spot represent one bit. A character is recorded across a section of the tape in eight bits. An extra bit called the parity bit is used for error detection. Altogether there are nine rows (tracks or channels) across the tape. The combination of bits to represent characters depends on the character set being used. There are two popular characters sets in use: the EBCDIC (Extended Binary Coded Decimal Interchange Code) and ASC11 8 (American Standard Code for Information Interchange).

The tape is made of strong and lightweight plastic coated with magnetic oxide material. It is, typically half an inch wide and 2300 feet long. The amount of information it can hold depends on the recording density used, which in turn depends on the density of packing of the tape (i.e. the closeness of the columns of magnetised spots across the tape). Magnetic tape is one of the devices called sequential access devices. That is, records are written into it and retrieved from it one after the other.

Prompted by the computer, the tape drive (the mechanism that revolves the tape) begins to spin the tape, starting from zero and accelerating to the appropriate speed before any "read" or "write" operation can start. While the tape is accelerating, the read/write head positions itself above the tape. Data are recorded in groups or blocks. Typically, 80 columns hold one record. The blocking factor is the number of records grouped together in one block. There are gaps between the blocks of data (inter-block gaps), which allow the tape to accelerate to the correct speed before a data block is reached. They also allow the tape time to come to a complete halt when the operation ends.

The limitation with sequential devices is that it is not possible to access records quickly and randomly. If records are to be read from different sections of a tape, it is quite clumsy, to move back and forth to do so. Sequential devices are ideal for backing up information in a system for security purposes. They are also useful as an input and output device for batch-processing applications in. which large volumes of data are processed on each run. Magnetic tape also comes in cartridges of 3 x 5 inches. A tape can hold about 200 MB of data.

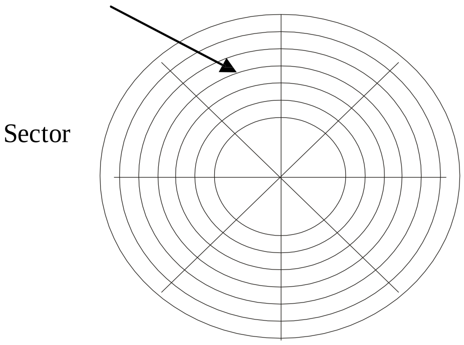
3.3 Floppy Disks

The ability to access and write individual records randomly on disk was a major advance in storage technology. It removed the need to process data entry and retrieval of records in batches. The invention of the floppy disk by Shugart Associates in 1972 followed the introduction of microcomputers. It proved to be a cheap medium.

A diskette is made of polyester material coated with magnetic oxide. There were two types of diskettes: the 5.25-inch type and the 3.5-inch type. The former is already out of the market. A 3.5-inch diskette holds 144 MB of data. Before the invention of the floppy disk, microcomputers used small cassette tapes. They had neither the speed nor the reliability needed for improving performance of computers.

A diskette is marked into concentric circles called tracks. The tracks are divided into sections called sectors. A new diskette has to be formatted, that is, marked into tracks and sectors. When a diskette is inserted into the drive, a pair of rings grips it, and it begins to spin. It accelerates up to the required speed for read for write operation (about 300 revolutions per minute). Meanwhile the read/ write head has moved to the right track to wait for the right sector. When that sector comes up, it does the reading or writing of a record or records. The time required to move the read/write head to the correct track is called seek time. On the track, the time for the desired record to rotate to the read/write head is called rotational delay. The time taken from finding the right record to reading it is called data transfer time; that is, the time it takes for the whole of the desired record or records to pass under the read/write head. The access time is the total time taken, including seek time, rotational delay and data transfer time.

Recording surface



SELF-ASSESSMENT EXERCISE

Get a bad diskette. Pull back the silver white shutter so that the thin magnetic medium is exposed. Take a close look at it and note its appearance.

34 Winchester Disks

The Winchester hard disk came out in the 1970s as a major success in magnetic storage device. It is product of highly sophisticated technology. Manufactured under "clean room" condition, the unit is permanently sealed in pressurised plastic cartridge. This excludes impurities like smoke and dust particles that could damage the disk surface or the read/write heads. Inside the unit, the read/write heads float above the disk surfaces in a cushion of air. This makes hard disk immune to friction and mechanical damage in the course of use.

The Winchester disk is much faster than the floppy disk. It starts to revolve as soon as tin, computer is switched on and continues to rotate as long as the computer is on, unlike the diskette, which starts spinning only when it is inserted into the drive. The Winchester disk rotates much faster too. Rotation is at 3600 revolutions per minute. It also has much greater capacity. By the year 2001 the capacity had reached 30GB. That is 30,000 millions of characters.

3.5 Optical Media

The invention of optical storage media is presently the greatest height that has been reached in storage media technology. Optical media are based on the properties of light. The first storage medium to come out of this technology was the compact disc audio player, which came into the market in 1983. The compact disc read only memory (CD-ROM) came next in 1985. Subsequently, the compact disk interactive and compact disc video came into the market.

The first optical disks that were commercially available were used on television videodisk systems. At that time each videodisk could hold 53000 picture frames and that gave half an hour of program or movie. A double-sided videodisk gave twice that. A special feature of optical media is that the laser beam can jump about, scan and play flames in any sequence under the control of a computer. This makes it especially suitable for storage for interactive applications.

The CD-ROM has been of tremendous importance for information storage and dissemination. Its durability, huge storage capacity, and high fidelity make it useful for distributing information in large volumes and for keeping materials of high archival value. Another benefit of CD-ROM is that it has made large databases available to individuals using personal computers. This is very important for: users in developing countries and for many libraries that had stopped subscribing to some major bibliographic databases in, the printed version on account of the huge cost that was involved. Regrettably they were not in a position to subscribe to the online service because of technical difficulties and prohibitive cost of telecommunications charges. They have now turned to the CD-ROM versions of those databases.

CD-ROM disk is a metal-coated clear plastic platter onto which a laser light burns digital information. Bits are stored on the optical disk by burning microscopic pits into the surface of the disk. The absence or presence of a pit determines whether a bit is zero or one. Laser is used to read and retrieve information from it. Data coding on CD-ROM is standardised. Manufacturers follow well-defined standards, so that the same CD-ROM can be used on different machines. A CD-ROM is usually marked into blocks of 2,352 bytes. Each block can hold 2048 bytes of user data, while the remaining bytes store identification information for the block, error correcting codes, and data concerning the mode of, storage. A single CD-ROM of 3.75 inches in diameter can hold at least 700 MB of data.

3.6 Micrographics

Micrographic media were introduced to deal with the vast storage space requirement associated with printed output. Micrographics are microfilms carrying information that has been recorded not in the, binary forms of magnetic and optical storage but in the actual characters of text and actual form of graphic images. The techniques of microfilming make it possible to reduce a page photographically to a very small fraction of the original size. This is where the advantage of reduction of storage space requirement comes in. However, there is the limitation that special equipment (reader) is needed to read the information in it. The microfilm reader magnifies the small images to be legible to the human eye.

Before the advent of optical storage media, microfilms were the best media for distributing voluminous and archival materials. They are quite cheap and portable. They also have long life if stored properly. Another important advantage is that the output of a computer operation can be stored, directly in microfilms and therefore bypassing paper output. In this case a COM (Computer Output on Microfilm) recorder is required to do the recording. Microfilm came in several formats.

Roll film: Roll microfilm has been the most widely used and perhaps the most economical in terms of the vast amount of information that it can store at minimal cost. Microfilms are commonly available in 16 millimetres and for special applications in 35 and 70 millimetres. One roll film can hold several thousands of pages of a book.

Cartridge microfilm: A second format is the cartridge microfilm. It has the same features as the roll microfilm, but it is more suitable where rapid document retrieval is required. It is used with a 3M Reader/Printer and allows any image to be located quickly and read, or if necessary, hard copies produced within seconds. Cartridges are usually filed in shallow drawers with each cartridge carrying identification information for easy retrieval.

Microfiche: A third format is called a microfiche. It is a rectangular negative with many rows of frames, images that correspond to pages of a document. It is read with the aid of a microfiche reader, and hard copies can be produced easily.

4.0 CONCLUSION

In this unit the storage devices that are commonly used were presented to you. You have also learnt how the computer represents data. You should now be able to describe the technical features of a number of storage devices.

5.0 SUMMARY

Now that you have learnt how the computer represents data and you are familiar with the technical features of various storage devices, you are in a position to move on to consideration of records and files, which are the subjects of the next unit.

6.0 TUTOR-MARKED ASSESSMENT

For each of the devices that were covered in this unit, list the special features that make it attractive as, a storage medium.

7.0 REFERENCES/FURTHER READING

Cyril, H. P. *et al*. (1982). *Information Systems Design*. Sydney: Pretence-Hall.

French, C. S. (1996). *Computer Science*. (5th ed.). London: Letts Educational.

UNIT 3 RECORDS AND FILES

CONTENTS

1.0 INTRODUCTION

In the last unit, you learnt how the computer represents data and you became familiar with the technical features of various storage devices. In this unit, you are going to learn about records and files.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- distinguish between records, files and directories (or folders)
- identify the elements of a bibliographic record
- describe the ways files are organised
- explain how the computer accesses files and records.

3.0 MAIN CONTENT

3.1 Personnel Records

When an officer talks about the record of a transaction, what exactly does he mean? He is probably referring to a letter or any of the documents that might include a receipt, invoice, a voucher, some quotations, a proposal and so forth. What makes a record in computer storage is quite different. The examples above qualify as documents in their own right and so must be stored as files.

Similarly, when a clerk brings out a file, you would expect it to contain documents that include memoranda, letters, financial statements, reports, receipts and so forth. What the clerk calls a file here is more appropriately called a folder or a directory in computer language, while the documents are files.

From now on you need to hold on to this new understanding. Now, let us visit the personnel or administrative department of a company to see what information it keeps on staff. Each employee is identified by name and personnel (or file) number. Other pieces of information include sex, date of birth, date of assumption of duty, designation, salary scale, and so forth. If this kind of information is to be stored in the computer and for all the staff, we have to consider how to arrange the pieces of information on each person. We have talked much about classifying and ordering things or information to facilitate access and retrieval. The pieces of information enumerated above will be stored in their own "compartments". In computer language these are referred to as "fields", while all the pieces of information for each employee make a record. All the records of all the employees make up a file.

SELF-ASSESSMENT EXERCISE

What is a file, a record, a field?

3.2 Bibliographic Records

A bibliographic record is made up of distinct pieces of information about an information source. Let us take as our example a book.

Carol Omu. *A Resource Book in Town Planning*. Lagos: Generation Press, 1998.

The pieces of information or fields are as follows:

Author	Carol Omu
Title:	A Resource Book in Town Planning
Place of Publication:	Lagos
Publisher:	New Generation Press
Year of Publication:	1998

For a journal publication such as

Low T. Jam. *Rice Production in Tropical Highlands*. Journal of Agronomy an Crop Protections. Vol.8, No.2, 1997, 23-35.

We have the following fields:

Author:	Low T. Tam
Title:	Rice Production in Tropical Highlands
Journal:	Journal of Agronomy and Plant Protection
Volume:	8
Number	2
Year:	1997
Pages:	23 - 35

If we can recognise all the fields in the bibliographic records of information sources, arrange a format for storing the records. We shall demonstrate this later.

3.3 Some Characteristics of Records

All the fields of a record contain useful information but they are not equally important for the purpose of identifying the person, thing or place that the record is about. For instance, it is more meaningful asking for the details of a book by the author's name or by the title than using the place of publication. Similarly it would make more sense to request the personal details of a staff in an office by his or her name than by his or her date of birth. In these two examples, the fields that are most suitable for requesting information about a book or about a staff are also the ones to use to get to their records in a computer file. They are called the primary keys. They are the best keys for getting at specific records in the files. Other fields could still be used as keys anyway but their identification value is only secondary. So they are called secondary keys.

The number of fields in a record could be fixed or could vary. For instance the record of a book that was written by two authors would not have the same number of fields as one that has one author, unless we want to lump all the authors in one field. That itself creates a problem in retrieval by each author's name. The fields themselves are not always of the same size. In our two examples of records, the field "name" and "title" would vary in size from one record to another according to the length of the name and title in each record. Fixed-length records are those with the same number of fields and the same number of characters in a field for all the records in the file. Variable-length records are those that have variable number of fields and variable field size.

3.4 Access to Files

It must be emphasised here that records and files are created and stored for a purpose, namely, for retrieval. The question to worry about here is how to locate a specific file and the individual records in it. You will recall that among the secondary storage media are magnetic tapes, magnetic disks and optical media. When files are stored in a sequential device, records are read one after another until the computer gets to the end of the device. So when it locates a file, it can search for the desired record until it finds it, or if the record does not exist, until it gets to the end of the device.

In the other devices such as disks and optical media, the computer can locate directly any file or record without reading the ones before to get there. Such devices are called direct access storage devices. With these devices the computer maintains a table of content, catalogue, or index of the name and location of each file in the device. Once the file is located the next thing is to locate the desired record.

3.5 File Organisation

File organisation is the way records are arranged and stored in a file. When the computer starts the process of storing (that is, writing) a record, the mechanism that does the writing called read/write head (it does the reading too) moves to locate the track and then the sector in the case of a direct access device.

The record to be written is sent from the main memory to the read/write head, which then writes it on the surface of the storage device. When the storage medium is a random access device, the read/write head moves rapidly to any location that is available to write on. The next record follows and is written in the next available position. On a magnetic tape, on the other hand, the read/write head moves sequentially from the beginning until it gets to the first writable section (if the tape is empty) or from record to record until it gets to the end of the last record (if some records are already in it). Then, it begins writing the record. The records that follow will be written on the next successive magnetic positions.

In the process of reading these records the computer follows exactly the same procedure to locate them and then the read/write head will transfer them to the main memory.

3.6 Disk Pack and Cylinders

There is only one recording surface on a single sided floppy disk and two recording surfaces on a double-sided floppy disk. Other disk assemblies (units) consist of a number of disks connected one above the other by a common drive mechanism. That gives 2n-2 recording surface, where n is the number of disks. The top and bottom surfaces are not used for recording. Each writable surface has its own read/write head. Each surface is marked into concentric tracks (about 200) and each track into sectors. With 200 concentric tracks on each surface, it would look like there are 200 concentric cylinders in the unit.

Fig. 1: A Disk Pack Showing Cylinder

3.7 Address

The special feature of the direct access storage devices is that any record can be located independently disk pack is given in terms of the cylinder number, the track number, and the block number.

3.8 Sequential Files

In a sequential file, records are written or read in a fixed sequence. Sometimes, the records have to be sorted by a key field before-they are written to the storage device. If, for instance, the staff records of an organisation are sorted by surname, then the whole records will be stored in an alphabetical order, of surname. We noted earlier that magnetic tape is a sequential storage device. That is, records and files are stored in magnetic tape in sequential order. They are also read in sequential order.

3.9 Direct Access Files

In the direct access storage devices, every record has an address that makes it possible to locate it independently of other records. A common practice is to maintain an index or table of the key and the relative record number of the record in storage. The index is always stored in sequence and this facilitates easy reference for the purpose of reading the records. To retrieve a record, the actual key is looked up in the index, and the relative record number of the record that matches the key is found. Then the address in storage is worked out and the recorded accessed.

3.10 Indexed Sequential Files

Indexed sequential file organisation is based on sorting of records by a key and the creation of cylinder and track indexes. First the records are sorted by an appropriate key and then written on successive locations on the disk. As a track is filled, the key of the last record on that track is entered in a track index. The next records go to the next track, that is the corresponding track below in that cylinder. When that is filled the last record also goes into the track index.

The process continues until all the tracks in a cylinder are filled. The last record in the cylinder is entered into the cylinder index. The cylinder index will continue to grow as the cylinders are filled. For each cylinder there will be a track index but there will be only one cylinder index for the whole disk pack.

Now, one can imagine how easy it will be to locate and retrieve records. The first thing is to look up the cylinder index and get the cylinder that bears the record requested. The next step is to look up the track index of that cylinder to find the right track. Having found the track, the record can be located by scanning sequentially.

SELF-ASSESSMENT EXERCISE

What are the main issues in file organisation and access?

4.0 CONCLUSION

In this unit you have learnt a number of things about records and files. You can now distinguish between records, files and directories (or folders), identify the elements of a bibliographic record, describe the ways files are organised and explain how the computer accesses files and records. File organisation is extremely important for the purpose of retrieval.

5.0 SUMMARY

You have now learnt how the computer represents data and you have become familiar with the technical features of various storage devices. You have also learnt about records and files and the ways that the computer accesses records and files. In the next unit, you will be introduced to databases. You will learn that databases are made up of files organised in such a way to facilitate retrieval.

6.0 TUTOR-MARKED ASSIGNMENT

Visit the computer department in at least three big organisations and find out what kind of storage devices they are using and in what circumstances they are using the following methods of file organisation: sequential, direct access, indexed sequential. Then write a report on file organisation and access in the organisations. Your report should be between four and six pages of A4 typed with double spacing in 12 points of Times Roman.

7.0 REFERENCES/FURTHER READING

Cyril, H. P. *et al*. (1982). *Information Systems Design*. Sydney: Pretence-Hall.

French, C. S. (1996). *Computer Science*. (5th ed.). London: Letts Educational.

UNIT 4 DATABASES

CONTENTS

1.0 INTRODUCTION

In the last unit, you learnt about records and files and how the computer stores and accesses them. In this unit, you will learn how files are organised into databases. Understanding of the structure or organisation of databases is important for the purposes of storage and retrieval of information.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain the problems of customised file approach
- explain the basic concept of database
- enumerate the advantages of centralised database approach
- describe the various approaches to database organisation.

3.0 MAIN CONTENT

3.1 Management of Data

Data and information are a major resource in an organisation. Management of data and information is, therefore an important business activity in every organisation. Go into any office and you will not fail to notice the movement of "files". In the traditional office, every unit or department had its own set of files. The personnel department had a file for each employee. The accounts department also had a file for each employee. In a situation which a file had to leave one unit to another, there was always a problem of misgiving, resistance or hostility.

Now, many organisations are making intensive use of computers. There is at least one computer, in every department or unit, and the records that were in cardboard folders (or file jackets) are now in the computer. What happened in most cases was that each department developed a customised computer based system of managing its own records. As you can recall, the computer requires a program or a suite of programs (software) to execute a job or perform a task. So according to the tasks performed in a particular department, that department purchased and installed the hardware, and developed (or purchased) and installed the necessary software, to manage its records. Before we examine some examples of these let us define the term "application".

An application is the software developed to perform a job comprising a set of related tasks. A large bookshop would be expected to have personnel file, inventory file, customer file, general ledger file, and payroll records, among others. In a situation in which each department or unit is doing its own thing, we would have the following applications:

Unit	**Application**	**Type of Data**
Personnel	Personnel	Name, Address, Designation, Birth, Marital Status, etc
Accounts	Payroll	Name, Address, Bank, Salary, Deductions, Designation, Allowances, etc
Accounts	General Ledger	Debits and Credits, Status of each account, etc.
Marketing	Customer	Name, Address, Credit information, Terms of contact, etc.
Marketing	Inventory	Price, Cost, Locations of items, Stock levels etc.

3.2 Problems of Customised File Approach

The problems with this customised file approach are many. To start with, there is the likelihood of data redundancy. The same data are replicated in several files. For instance, in the table above, some personal details of each member of staff are duplicated in the personnel and accounts records. Now, think of a situation in which an employee changed his address and communicated this change to the establishment. The personnel unit updated the address of the employee in its records, but through somebody's negligence or through communication gap, the accounts unit was not aware of the change of address. The consequence is that the personnel unit will be using the person's current address while the accounts unit will continue to use the old address. This is a case of lack of data integrity. That is, there are different versions of a piece of information about the same person or thing.

Another problem is the tendency for a unit to hold on to what it considers its own. The data generated and the files treated by a unit are usually protected by that unit. There is the problem of ownership. No unit feel obliged to share its data with other units. With the problem of data ownership comes the problem of accessibility. The control of data by one unit makes it difficult for others to have access to them.

There is another problem you may need a lot of imagination to grasp. That is the problem of data dependent program or data dependency. When computer applications are implemented in different departments in an isolated fashion, the way the data are defined to the software is different. Let us consider the example of date. For one application, the date may have been defined as dd-mm-yyyy, while in another the format was given as ddmmyyyy. In the first format, there are two spaces for day then a space for the dash, then two spaces for month, then a space for the dash, and lastly four spaces for year. The computer would then allocate ten spaces for the date field. In the second case, it will allocate the first two spaces for day, the next two spaces for month and then four spaces for year with no spaces between them. Then there would be eight spaces for date. This means that when data have been defined differently to different applications, data meant for one application cannot be used by another one. This is a case of software being locked to its data and data being locked to the software for which they were stored.

SELF-ASSESSMENT EXERCISE

What are the problems of the customised file approach?

3.3 Database Concept

The problems described above can be solved only through centralised planning and systems Development. The various files and records could then be integrated into a database. A database is a collection of data or data files that have been integrated in a manner that facilitates easy access and haring. Once the format of storing data has been standardised for the whole enterprise every program an access the data without problem. The centralised control of the database eliminates the problems of ownership and accessibility. Another point is that since there is only one database for everybody, there is', no longer the possibility of either data redundancy or lack of data integrity.

SELF-ASSESSMENT EXERCISE

Now, can you recall the advantages of having a centrally controlled database as described above?

3.4 Advantages of Centralised Database

1 Data redundancy can be minimised

We would actually say that data redundancy is eliminated. However, there are instances when some level of redundancy is needed. The important thing is that unnecessary redundancy can be avoided. Another way to put it is that redundancy is controlled.

2. Consistency and integrity of data can be maintained

A situation in which there is conflicting information coming from different departments or units is quite objectionable. When the same data elements are scattered or replicated in several units, there is no assurance that if they are updated in one unit, they will be updated in all units. When there is only one entry of a data element, then the problem of inconsistency cannot arise. This means that it will be possible especially with the aid of the software that maintains the database to ensure accuracy of data, that is data integrity.

3. Sharing of data becomes easier

The question of any unit trying to hold on to its data does not arise with a centralised database. All the applications that need to access the database can do so without difficulty.

4. Data independence can be maintained

Some software packages dictate the way data should be stored for their use. Unless data are presented that way they cannot read them. Also it is the way data have been declared to software that it will expect them. That is, a software package could have difficulty accessing the fields in a file created for other application. This is the problem of data dependency. This problem is solved in a centralised database because such a database is managed by software that sits between the database and the various applications that access the database. It serves as an intermediary. Even if the characteristics of a field should be changed it is only that software that needs to know about it. The applications do not need to know about the physical storage of data elements. This is data independence.

5. Standards can be enforced

Centralised planning ensures not only uniformity in the construction of data structures (the format of storing data elements) but that all necessary standards are met. It is the standardisation of the format of data storage that allows different programs to access the database.

6 Conflicting needs can be reconciled

The process of building a database is that of consensus building. Every section has to give way to what is best for the whole organisation. It could mean a complete overhauling of the entire organisation for better performance.

7. Security controls can be applied

When the data resources of an organisation are centralised to be accessible to all sections, there is security problem. This is probably one reason why people may resist the idea of centralised control. There is really no cause for alarm as a database system normally has a number of security devices. The first security device is the password, to ensure that only authorised persons can get into the database at all. Secondly, different authority levels are assigned to those who are authorised to get into the database, to define what fields or data elements they can access and what they can do. Somebody who has no business with a particular field will not be given access to it. Someone else may have access to it but he can only retrieve information from it. He cannot change the content of that filed.

8. Valuable information is received

Information that is supplied to management is more valuable since it is based on a comprehensive and integrated collection of all the data files of the organisation instead of drawing from disparate files.

9. Various reports are available

Besides routine reports, all kinds of ad hoc reports are made in response to various requirements.

10 There is economic gain

The fact that data are not duplicated amounts to economic gain. The gain is further enhanced through the minimised input preparation, as there is only one input operation.

11 Programming time is saved

Considerable saving in programming time is made because the software that manages the database does creation and processing of files and the retrieval of data.

SELF-ASSESSMENT EXERCISE

Can you now sum up what makes a database?

3.5 Database Organisation

Four approaches may be adopted for organising databases, according to which we following types of database.

- Relational
- Hierarchical
- Network
- Distributed

Relational databases

Table 1: Relational Database

Name	**Matric No.**	**Date of Birth**	**Faculty**	**Department**	**Year of admission**
Bade, J.J	128858	12-09-	Arts	English	1997/98
Durojie,L	232359	04-08-1982	Science	Geology	1998/99
Kabogu, G	265583	23-12-	Science	Chemistry	1998/98
Adeyi, M	139750	09-04-1971	Agriculture	Agronomy	1997/98
Zaki, O	375838	07-02-1984	Soc	Economics	12000/01

In relational database, files are organised in tables called relations. Records are arranged in rows while the fields are arranged in columns. In technical jargons, the rows are called tuples and the columns attributes. Relational databases have already become the most common type of database. A relational database is accompanied with index files to facilitate access.

Hierarchical databases

A hierarchical database has a tree-like structure. The organisation has the advantage of efficient use of storage space and rapid access. The database has a top-level record type known as the root node. All access operations to the database must start at the root node. The root node is connected to lower-level nodes, which contain data elements that are attributes of the root node. These lower nodes are in turn connected to yet lower level nodes.

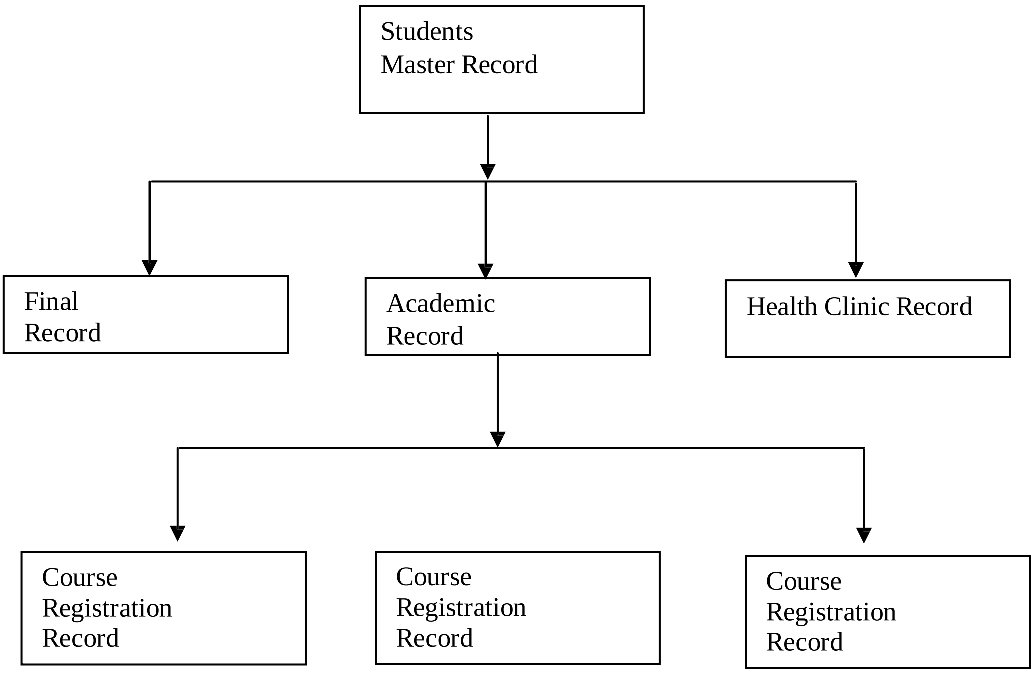


Fig. 2: Hierarchical Database of Students Record

Network databases

Network databases are based on the same principles as the hierarchical databases. The network model makes it possible to develop many to many links between fields. There is no single root node. Instead, access can be made through any level in the network and one can move upwards or downwards.

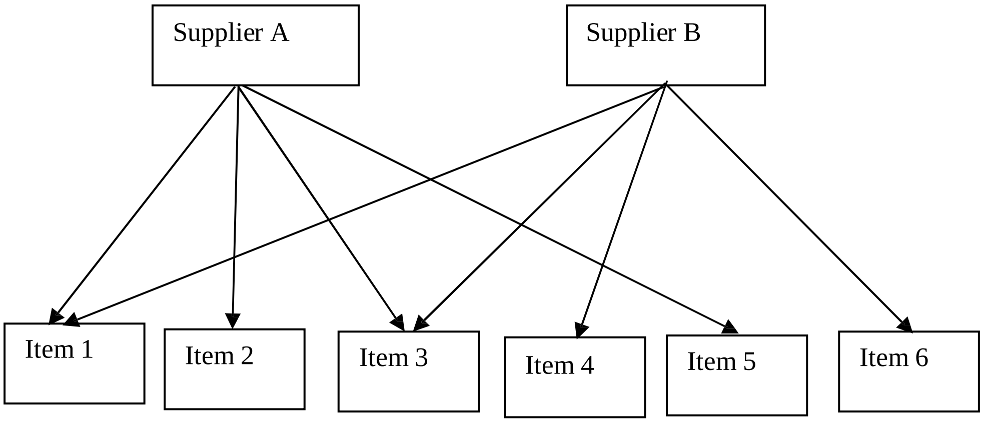


Fig. 3: Network Databases

Distributed databases

With progress in database management software and telecommunication links, many organisations no longer need to store all of their databases in one physical location. Instead, they are spread across a network of computers that are dispersed in widely separated geographical areas and linked through - telecommunication facilities such as telephone lines. Such a dispersed database is referred to as a distributed database. A good example is when a bank maintains a database of its customers at each branch, which can be accessed not only at that branch but also from other branches. A network of bibliographic databases could be seen as a kind of distributed database. A user who logs into anyone of them can retrieve information that is stored in any other one. A distributed database appears to a user as if the entire database is stored at the location from where he or she is accessing it.

4.0 CONCLUSION

In this unit you have learnt a number of things about databases, especially how files are organised into databases. You have also learnt four approaches to database organisation. You should now be able explain the basic concept of database, the problems of customised file approach, and the advantages of centralised database approach.

5.0 SUMMARY

Understanding of the structure or organisation of databases is important for the purposes of storage and retrieval of information. With all that you have learnt about databases in this unit, you are now ready to learn about the software that manages databases in a computer.

6.0 TUTOR-MARKED ASSIGNMENT

Write an essay on "Databases: The Key to Information Storage and Retrieval.” Your write-up should be between four and six pages of A4, typed double-spaced with 12 points Times Roman.

7.0 REFERENCES/FURTHER READING

Cyril, H. P. *et al*. (1982). *Information Systems Design*. Sydney: Pretence-Hall.

French, C. S. (1996). *Computer Science*. (5th ed.). London: Letts Educational.

UNIT 5 DATABASE MANAGEMENT SYSTEM

CONTENTS

1.0 INTRODUCTION

In the last unit, you learnt the fundamentals of databases. You will now see how the computer manages databases with the aid of a database management system, which also serves as an intermediary between the physical data storage and the various applications that have to access the databases.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain what a database management system is
- explain the concept of database schema
- explain the concept of query languages
- describe file management system
- describe transaction processing monitors.

3.0 MAIN CONTENT

3.1 What is a Database Management System?

The business of managing a database and giving access to applications programs is undertaken by the software called Database Management System (DBMS). Database Management System has been defined as the software which organises the structure of the database (through the data-description language or DDL) and handles all access to the database (through the data-manipulation language or DML). It is the DBMS that sits between a database and application programs that need to access the database. It acts as an intermediary. The DBMS facilitates the access to and updating of the database. Since it provides standard interface (link) to the data, it provides a means of implementing tight and consistent integrity control. The advantages in the centralised database approach to data management, which we enumerated earlier, are possible because of the management capabilities of the DBMS.

The DBMS has to meet the needs of three categories of users. First are the end-users who are usually not programmers. The DBMS has to provide an easy interface to access the database. It is expected that the DBMS provides a high level of user-friendliness through dialogue boxes, error messages, and help facilities. The second category of users comprises programmers and system designers, who have to be pre-occupied with the technical details of efficient access to and manipulation of the database. Lastly, the DBMS has to serve the needs of the database administrator. It is the job of the database administrator to handle all issues about database design and definition, standardisation, database creation, access control, database maintenance, and security.

3.2 Database Schemas

Earlier, we talked about data independence as one of the benefits of implementing a centralised database. The DBMS shields application programs from the details of the physical storage of data. Rather it offers them a logical view of data. In other words, beyond the requirement of data independence, the DBMS makes it possible to present a logical definition of the database to users and their application programs. This logical definition is called a database schema. It defines the record types, fields, the type of values a field may have, repeating groups, and relationship between records. Thus there are two views or models of the database, namely the physical model and the logical model. The physical model describes such characteristics as access paths in the database, the linkage between records, the formula used to calculate addresses, the means used to handle overflows, the representation used for numbers, as well as other physical aspects of storing, finding, and retrieving records.

Each application program that accesses the database also has its own view or schema of the database, which can only be a subset of the entire database. Such a view is called a subschema.

SELF-ASSESSMENT EXERCISE

Now, consider this. The database of the staff of a company is made up of records with the following fields: name, employment number, date of birth, rank, salary scale, tax code, deductions, gross salary, net salary, section, housing allowance, transport allowance, total allowance, sex, marital status, residential address, home address, next of kin, religion, date of employment, whether appointment has been regularised, and bank.

Can you guess what fields will be included in the sub-schema used by the payroll in the Accounts Section?

3.3 Query Languages

The DBMS mediates between application programs and the database. That means that it has to receive a request or a query from an application program each time the program has to access the database. Such a query is given in what is called a query language. A query language usually comprises a DDL and a DML. The DDL is used, to specify the data in the database while the DML is used to access the data. DDLs and DMLs may by independent of any other language or may be embedded in another language. The combination of a DIAL and a DML is called Data Sublanguage (DSL). An example of a DSL is the Standard Query Language (SQL). The diagram below shows the process of communicating with the DBMS. The user's applications programs are all connected to the DBMS server. The DBMS server is a special program, which can accept multiple connections from many applications programs at the same time (a capability referred to as multi-treading).

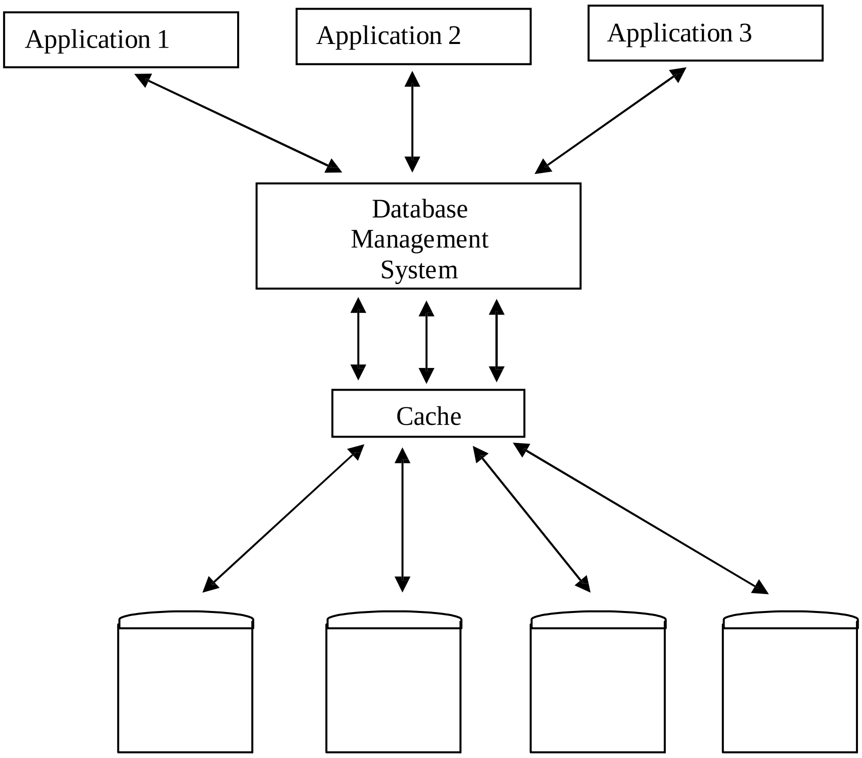


Fig. 1: A Database Management System

A request is initiated by an application program and sent to the DBMS server using a DSL commonly a SQL. The DBMS translates the request and recognises it to be one of many possibilities such as a request to define, insert, update, delete or retrieve specific data. If for instance, it is a request to retrieve data, the DBMS server will access and retrieve the data requested and puts them in a data cache from where they can be more rapidly manipulated.

Query languages operate in two basic modes, namely terminal monitor mode and as embedded query languages. The terminal monitor mode allows a user to use the query language at a terminal. It allows the end-user to formulate queries to receive information from the database. In the second mode, the query language statements are incorporated into the program codes that have been written for other languages like COBOL and C. The SQL is one of the most widely used query languages, and has become an international - standard for database query languages. Among the many computer manufacturers and database product suppliers that have adopted it are Digital, IBM, INGRES, ORACZE, INFORMIX. It uses plain English verbs to request for action, for instance, CREATE, SELECT, SET, INSERT, UPDATE, DELETE.

3.4 File Management System

One of the greatest developments in the history of computers and information management is the success of microcomputers and their popularisation. Part of the success story is the proliferation of software intended to run in them. Microcomputer versions of most software packages that were meant for mainframe and mini-computers have been developed. In recent years a wide range of software packages with the features of a database management system have come into the market. They are actually referred to as file management systems.

Usually they have rudimentary DDLs and DMLs, which make it possible for users to set up and maintain a number of files without much programming effort. They are quite efficient in their retrieval function as in well such tasks as selecting data and sorting. They are based on relational database approach. Some of them have become so well developed that they can be regarded as real database management systems. The DBMS are quite simple since they are not required to serve a large number of users. One factor that makes them quite attractive is that they are easy to use and easy to programme. Examples of these products are Microsoft Access, dBase, Dataease, INGRES, FoxPro, Paradox and ORACLE.

3.5 Transaction Processing Monitors

We should mention here the highly specialised programs called transaction processing monitors (TP Monitors), which perform transactions for applications. They are not anything like DBMS. Our interest here is that they are often used in conjunction with DBMS to pass requests from applications to the, DBMS server and transmit the response back.

The benefit of having the TP Monitor to assume this role is that it queues up the requests from the applications and processes them one after another. Thus it reduces the number of connections from applications programs that the DBMS server has to manage. This is really an important relief in a situation when there is a large number of applications that are all repeatedly performing a small set of similar transactions.

4.0 CONCLUSION

You have now completed the fundamentals of database management systems and you should be able to explain what a database management system is, the concept of database schema, and the concept of query languages. You should also be able to describe file management system and transaction processing monitors.

5.0 SUMMARY

You have now seen how the computer manages databases with the aid of a database management system, which also serves as an intermediary between the physical data storage and the various applications that have to access the databases. In the next unit, you will begin to explore the concept of information retrieval system.

6.0 TUTOR-MARKED ASSIGNMENT

Describe the role of a database management system in information storage and retrieval. Your write-up should be between four and six pages of A4, typed double-spaced with 12 points Times Roman.

7.0 REFERENCES/FURTHER READING

Cyril, H. P. *et al*. (1982). *Information Systems Design*. Sydney: Pretence-Hall.

French, C. S. (1996). *Computer Science*. (5th ed.). London: Letts Educational.

MODULE 3

UNIT 1 CONCEPT OF INFORMATION RETRIEVAL SYSTEM

CONTENTS

1.0 INTRODUCTION

In this unit, you will find out exactly what a retrieval system is all about. We shall begin by considering what it means to retrieve, information.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- describe a document retrieval system
- explain the concept of information retrieval
- describe the organisation of an information retrieval system.

3.0 MAIN CONTENT

3.1 Document Retrieval System

We shall begin our consideration of what an information retrieval system is by describing a document retrieval system. Let us start with a visit to a library. If you are not yet familiar with the functions of a library, it is advisable that you actually go and move around in one nearest to you.

As you enter the library, you are right at the security desk. The security officer asks for your library card. You do not have any because you are coming to the library for the first time. He directs you to someone at the circulation desk for your registration. Some forms are given to you to fill and you fill them right away. The stuff inspects them, and finds them in order. He nods in satisfaction and fills out a card, which he slips into your hand. That is your library card. He also gives you four small cards and says, "These are your borrower cards". This library is a small one and there are no computers here.

You want to begin your tour from the section called Acquisitions. You are shown a door at one end of the hall. On going through it, you meet a staff who is the acquisitions librarian. On several low shelves are books that have just arrived. They were purchased from different sources. They are being recorded and stamped. You are told that they will later be moved to the cataloguing section.

Now you walk through a side door and you are in the cataloguing section. Several persons are working on the materials that had recently been brought from the Acquisitions. In this section, they determine the subject area of each material. If, for example, a book is in history, they find what aspect of history it deals with. So, for every material, they note the subject and the aspect of the subject. That is called classification. They also have a way of describing the materials in terms of author, title, publisher, and place and date of publication. These pieces of information about the material are typed on cards. The process of describing the book is called cataloguing. On the catalogue cards, they indicate the subject and a code called class mark. The class mark represents the subject and aspect of the subject as well as the location where the material will be kept on the shelf. You can see a carton of books that have been processed being carried away for shelving.

The next section to visit is the circulation section. Now you are back at the circulation desk. Some people are there to borrow books. Others are there to return books. After watching for a while, you decide that you will also take a book away. You want a book that will introduce you to computers. You are told that you cannot just go about looking for a book in the library. There is a procedure to follow. You first search in the catalogue. Since you do not have any particular book in mind, you will have to go to the subject catalogue. Now you have to inspect the cards that represent books on various aspects of computers. After flipping over several cards you see the one that reads "Computers, Introduction to the Use". That is the one you want. The author is B. M. Egbon. If you had known of this book and the author, you would have gone to search the author/ title catalogue. Anyway, you write the class mark and then with the class mark and following the direction of the circulation officer, you go and get the book from the shelf. With the book in your hand you return to the circulation desk where it is charged out to you to take home. Congratulations! You have successfully retrieved a document from a document retrieval system. The diagram below illustrates a document retrieval system.

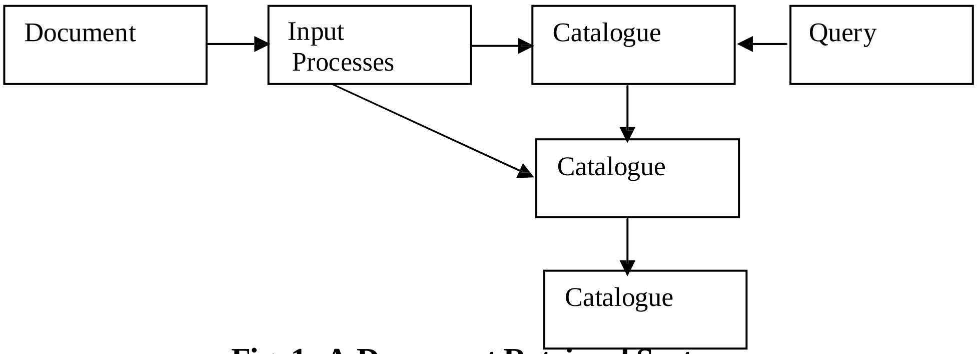


Fig. 1: A Document Retrieval System

The input to the system is a document. After it has been received, classified and catalogued, (that is after it has gone through all input processes), a card that will represent it is filed in the catalogue while the document itself is taken to the shelf where it will remain as in a document store. There is something interesting about the catalogue. Every document must be represented in it. Each card that represents a document is in this sense a surrogate, as it stands in for a particular document.

A person who needs information comes in with a query. A query in this context is an expressed need for information or for a document. It must be formulated in terms that facilitate a search for the information or document. Your query was "Give me a book on Introduction to Computers." You took that query to the catalogue and there you retrieved a surrogate for a book entitled "Introduction to the Use of Computers" written by B.M. Egbon. That card or surrogate bears the bibliographic record or bibliographic information of the book it represents. So, this first level of retrieval is called bibliographic retrieval. A successful retrieval at this point assures you that there is a document that will meet your need. The final retrieval is for the physical document and the output of the whole retrieval exercise is a physical document.

At the rate at which the computer is transforming all human endeavours, the library operations even in small libraries ire being computerised. So, you will not always find the document retrieval, system to be a manual system. For instance, you may not see card catalogues in some libraries because they have been replaced with computerised ones.

SELF-ASSESSMENT EXERCISE

From what we have seen thus far, what would you consider to be the elements of a document retrieval system?

We can distinguish four elements in a document retrieval system. The first is the input. The input comprises the documents that have been acquired and have arrived. The input processes include the acquisition, the recording, classification and cataloguing. The second element comprises the files to be searched. In the example above, we have two places to search, namely the catalogue and the shelf, the third is made up of the searching methods. In the example above, a search can be made in the catalogue through the author, title or subject. In the shelf, search has to be made through the class mark. The fourth, element is the output of the retrieval effort, which in our example was a document. Of the four elements, the input receives disproportionately large amount of effort and attention in traditional (particularly manual) systems in order to create a system that is easy to use.

3.2 Information Retrieval

To understand the concept of information retrieval, let us go again to your library. Assume that this time you want to get the address of UNESCO. Your first point of call is the circulation desk. This time you are not looking for a book to read but to get an address. Somebody at the desk tells you to go and see the Reference Librarian. The Reference Librarian gives you a seat and asks you what you want. Your answer is simple, "To get the address of UNESCO." The gentleman excuses himself and goes among the shelves. Two minutes later he comes back with an international directory of organisations. He puts the book on the table before you, opens the pages between which he had put his finger, and shows you the address you want. You write it in your address book and you are satisfied. What has been retrieved for you is information.

3.3 Information Retrieval System

The diagram below illustrates an information retrieval system. The input may be a document, some data or information. The input process includes recording, subject analysis, and possibly some reorganisation, especially in case of data. The input processes generate the representations or surrogates to be stored in a database.

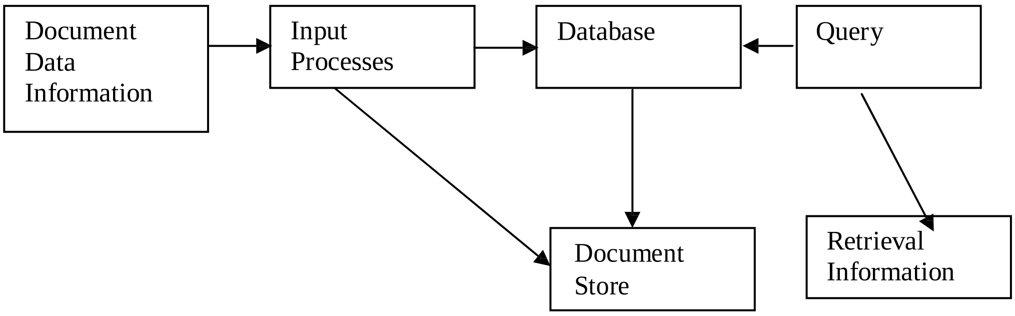


Fig. 2: An Information Retrieval System

What is retrieved will depend on the query. If the request is for some factual information, the procedure will begin from the database, move to the document store to find an appropriate document, and end with the delivery of the information requested. A second possibility is a request for one or two documents that will give information to help solve a problem on hand. There is still a third possibility that the request is for a list of references in an area of interest. In other words, the person who is making the request wants to know what information is available in that area. A researcher often makes this type of request. He or she would like to find out what work has been done and what has been published in the area of his or her research project. The immediate need is not a physical document. To satisfy this need, it is necessary to search the database. The output is bibliographic information.

SELF-ASSESSMENT EXERCISE

The information retrieval system shown in the diagram is actually an idealistic one. We have just, described what seems to be a complex or an all-inclusive type of information retrieval system. What kinds of information need can it respond to?

In real life not many libraries and information centres provide services to meet these needs with equal emphasis. For instance, most libraries and document supply organisations put more emphasis on delivering documents. Most of the organisations that offer information retrieval services actually concentrate on bibliographic information, sometimes with full text.

At this point we should try to distinguish between document retrieval system and information retrieval system. This distinction is not very easy to make. We know that a document retrieval system should deliver a document, an information material. It is the business of the person who gets it to read it and find or discover useful information in it. On the other hand, we expect an information retrieval system to deliver to a user the information he or she requests for. Now, what is the difference between a document and information? It is a fact that information is often carried in documents. If you ask for information on the last budget of the Federal Republic of Nigeria and someone gives you a newspaper or a pamphlet containing the budget speech of the President of the Federal Republic of Nigeria, you have a document all right, but in it you have the information you need. However, if the person consults the budget document and then gives you a rundown of the main items of the budget speech, he has given you information. In this case, you would quickly conclude that information retrieval has to go beyond document retrieval to offering information in the form (possibly in a synthesised and concise form) that the user can apply directly.

Of course you must remember that the kind of information people need varies. If what you want is not a document but a list of documents in the library or publications available in software marketing in Africa, the search for information would simply end at the library catalogue or in some database and you would have the information you need. So you can see that the distinction between a document retrieval system and an information retrieval system is not a matter of organisational structure but a matter of service emphasis. A system that provides documents in response to queries is a document retrieval system, while those that provide answers to specific questions, give some direction on where information can be obtained, or produces a list of sources of information are information retrieval systems. Modem information retrieval systems are computer-based systems.

4.0 CONCLUSION

You have now learnt what a document retrieval system and an information retrieval system are. You should now be able to describe a document retrieval system, explain the concept of information retrieval, and describe the organisation of an information retrieval system.

5.0 SUMMARY

In this unit, you have found out exactly what a retrieval system is all about. In the next unit, you will learn to appreciate the role of information centres in storage and retrieval.

6.0 TUTOR-MARKED ASSIGNMENT

Describe the organisation of a document retrieval system and an information retrieval system and explain how one is different from the other.

7.0 REFERENCE/FURTHER READING

Lancaster, F. W. (1979). *Information Retrieval Systems: Characteristics, Testing and Evaluation*. New York: Willey.

UNIT 2 THE ROLE OF INFORMATION CENTRES

CONTENTS

1.0 INTRODUCTION

You have just learnt about the organisation of retrieval systems. At this point we go into information retrieval, it would be necessary to consider the role of information centres in information storage and retrieval. This unit will present to you various information centres.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- define 'Information Centre'
- explain the role of various information centres in information storage and retrieval.

3.0 MAIN CONTENT

3.1 What is an Information Centre?

An information retrieval system cannot exist except in an organisational context. It is a functional part of an information centre. Information centres are the organisations that acquire, organise, store, retrieve and disseminate information. Libraries are a familiar example. There are other centres that do not give information directly but refer users to the tight sources. They are called referral centres. Some other, centres provide services in classification, cataloguing, indexing (subject analysis) and abstracting, and literature searching. There are also centres that keep track of current research. We shall examine each category and see in more details what each centre does.

3.2 Libraries

The very first function of a library is to acquire information sources, that is, information materials or documents. They include books, journals, reference sources (such as dictionaries, encyclopaedias, directories, yearbooks, manuals, bibliographies, sources that list the information materials which are available and which users can look for), indexes (sources with similar contents as bibliographies), abstracts (source containing brief reports on the contents of documents), theses and dissertations, maps, magazines, newspapers, government publications, reports from various institutions, other miscellaneous publications, and on-book materials, including audio-visual sources.

The next function is to classify the materials according to their broad subject and their specific subject area. In the process of cataloguing, the materials are described in terms of the bibliographic information (including author, title, publisher, place and date of publication, and so forth, in so far as they are applicable). These bibliographic details together with class marks (indicating the subject categories) are recorded on catalogue cards and filed in catalogue cabinets, or they are recorded with a computer and included in the computerised catalogue called database.

Another major function is the circulation of the library materials or user service. The information materials are meant to be used. Users are to receive as much help as possible to make maximum use of the materials. To retrieve information a user has to consult the catalogue to find out if the particular information source he wants is the library. If it is found in the catalogue (showering that the book is in the library) then it could be fetched from its storage location. Another user may need a comprehensive list of documents, while yet another may want factual information. The library should be able to meet all these needs. The library is also expected to go beyond providing information sources to users who come into the library. It should be able to send information to prospective users and members of their user community information selected according to their individual interests, before they are even aware that such information is available. This is called selective dissemination of information (SDI). Lastly, we should expect a library to be able to refer requests that it cannot satisfy to the information centres that can meet them (referral function) or in fact obtain from such centres the needed information materials on behalf of its users. There are many more functions that libraries carry out according to their types and their user communities; but here we are concerned with only those that are relevant to information storage and retrieval.

3.3 Referral Centres

Referral centres are the organisations that direct information seekers to sources of information or that direct people to where they can get what they want. When we were considering the functions of the library we said that it should also function as a referral centre. As a matter of fact every information centre should have this function. The referral function in this case is based on the principle that a request .that cannot be met by a particular centre should be referred to another centre that is a better position to meet it.

However, there are information centres that may be regarded strictly as referral centres. They are always busy attending to requests on where to find one thing or the other. Some of them are highly specialised, for instance, one specialising on employment information. Some organisations offer assistance in locating experts within national or regional boundaries and in specific fields.

SELF-ASSESSMENT EXERCISE

Can you suggest how information centres are able to offer referral services?

The input operation is usually elaborate and intensive as well for the kinds of services we have just mentioned to be possible. Take the case of an information centre that gives information on experts in selected fields. One way to build up its database is by conducting a survey. A questionnaire (a form for obtaining information from people) is sent through the institutions and agencies that employ such experts or that can reach them. The experts use the questionnaire to supply information on their background, including qualifications, specialities, the kind of jobs they are doing or have done, the organisations they have worked for, and so forth.

When the information centre receives the completed copies of the questionnaire, it will analyse them and create a record for each expert in its database. In response to a request, the database can be searched by name of expert for information on a particular expert, by area of specialisation in order to find out who are the experts in that area, as well as by other fields. Among the international organisations that would have built substantial database of experts are UNESCO, UNDP, and Commonwealth Secretariat. Maintenance of such databases is a continuous process. New records will have to be added while some are deleted.

Two other examples are industrial and educational referral centres. An industrial referral centre should have databases on industries, experts in the industries, products and services in the industries, sources of raw materials, imports and exports, and so forth. A referral system that students would be delighted to use is the type that answers such questions as, "What institutions have such and such a programme?" or "What financial aids are there and who offer them?" There are more questions that the system could answer. Comprehensive educational databases would be the basis for a referral service of the kind needed. The National Universities Commission in Abuja should develop such databases and create the needed services.

In real life, centres that are exclusively referral centres are rare. Many take up such a function as a secondary activity or at best combine it with some other functions.

3.4 Bibliographic Control Centres

The need for bibliographic control arose from the difficulty of coping with information explosion. The volume of publication has been increasing astronomically that no one can cope with more than a very tiny proportion of the information available in his area of interest. The question then, is "How does one get the most relevant information while ignoring bulk of the information, which is not really needed?" The individual information user needs help and some other people must supply that help. The latter have to keep track of all the information coming into existence and draw the attention of the user to it.

A lot of effort, skill and expenses go into the input process of such centres. It is not possible for any centre to take on the whole spectrum of knowledge. Bibliographic control centres have to acquire or have access to all the documents that have come out within their areas of mandate and within a specified period of time, carry out detailed subject analysis of the documents, and store the bibliographic information in their databases. Based on these databases, various services are offered by the centres. Usually, since centres concentrate on one subject area which may be broad or narrow, the services they offer are subject-oriented. Some others build up databases on specific problems and so offer, mission-oriented services. There are others that collect and analyse documents of all subjects and interests within national boundaries. They produce national bibliographies.

SELF-ASSESSMENT EXERCISE

What are the benefits of bibliographic control? What are the services provided?

3.5 Alerting Service Centres

Alerting services (also called current awareness services) are those intended to notify users of new information or new information materials on a timely basis.

The services include:

- Current contents
- Current bibliographies indexes
- Indexing and abstracting services
- Selective dissemination of information.

You will notice that three of these services were listed among the services offered by bibliographic control centres. The real consideration in alerting services is timeliness. Users must be informed of new information within a couple of weeks. One way to achieve this is to send out the content pages of issues of journals that have just been printed or are in press. Easy access to electronic versions, of journals will further improve the timeliness of current awareness.

3.6 Research-in-Progress Information Centres

Of special importance to researchers is information on the research efforts that are going on. Just as they, need to know what work has been completed and published so also they need to know of projects that are in progress to avoid taking on the same thing that some other person has started. Ideally, before a researcher finalises his research topic, he needs to carry out exhaustive literature search using bibliographic control sources and alerting services as well as services offering information on research-in-progress.

Information on research-in-progress is usually offered by agencies that have some responsibility for research administration. Sometimes such services are the initiative of documentation units of research departments or institutions. The input to their databases comes from questionnaire that the researchers fill and return. The information collected includes, Name of Organisation, Department, Title of Project, Duration, Expected Date of Completion, Name of Principal Investigator, Names of Collaborators, Funding Agency, and brief description of the project. The databases are updated on a regular basis. They provide the basis for practical information retrieval and a periodic publication. Users can request for information individually. There may also be served by selective dissemination of information.

3.7 Centres Providing other Services

There are a number of services that support information storage and retrieval, which are being offered by some centres. Among these are cataloguing and classification, indexing, technical literature distribution.

The Library of Congress in the United States has been playing a leading role in the classification and cataloguing of information materials and distribution of cataloguing information. Libraries all over the world that are using the Library of Congress Classification subscribe to the cataloguing information. Some centres specialise in indexing and providing indexing expertise and support. Some of them have developed thesauri, which are available to other centres. As you can recall, a thesaurus is a vocabulary of controlled indexing language, formally organised so that a priori relationships between concepts are -made explicit? The index language of a system with a controlled vocabulary is usually set out in lists, which may be in the form of subject headings, thesaurus, or some classification schedule. A thesaurus is meant to be a tool to aid the indexer in the choice of indexing terms for consistency and to assist the searcher in using the same terms as the indexer for maximum retrieval results.

Technical reports and translations are usually not available through normal book trade channels. Technical reports are usually the products of government-sponsored projects, private sector technical initiatives, projects executed by research institutes and investigations by international organisations. It requires painstaking efforts to track them and acquire them. In Nigeria there ought to be an agency or a unit in an agency that should be charged with systematic collection of such reports, analysing them for storage, and retrieving them for dissemination to the user community or delivery to individual users.

The business of translation is one that requires a lot of resources. A large proportion of information in any subject is in languages other than English, especially Russian, Spanish, French, Chinese, and German. Not many information centres translation of information sources from other languages. In a country like Nigeria, it would be advisable to create a translation agency to provide translations to researchers. A publication that is identified through bibliographic services to be relevant and important to research project should be ordered through such an agency. It would be the responsibility of the agency to acquire the publication, translate and store it, and provide a copy to the user who requested for it.

3.8 Information Analysis Centres

Information analysis centres function in the same way as do bibliographic control centres. However, information analysis centres add synthesis and evaluation to their input processes. Usually, a bibliographic control centre tends to collect documents exhaustively in its area of interest. An information analysis centre does not aim at exhaustive collection but at maximum quality of retrieval outputs.

Information analysis in this context involves evaluation and synthesis of information. Rigorous evaluation of information sources is done in the process of selection of material for processing. Subject analysis is done, and a synthesis of the text is undertaken. The centre responds to requests with searches, the outputs of which are critically evaluated before they are given to the user. This kind of service is not very common and is available for very specific research areas.

3.9 Document Supply Centres

There are some centres that specialise in document delivery. Usually they supply photocopies for a fee in response to requests for documents coming from both libraries and individuals. The British Document Supply Company (formerly known as the British Lending Library) is a good example of document supply centres. Many libraries find it much cheaper to use their services for obtaining journal articles from journals that are not highly demanded than to subscribe to such journals. Document delivery is being facilitated through online request and delivery especially through retrieval from CD-ROM.

SELF-ASSESSMENT EXERCISE

Now, try to summarise functions or benefits of these information services.

4.0 CONCLUSION

In this unit you, a number of categories of information centres were presented to you. You have learnt their roles in information, storage and retrieval and you should now be able to describe them.

5.0 SUMMARY

You have just learnt about the organisation of retrieval systems. You are now also aware of the role of information centres in information storage and retrieval. The next unit will introduce the issue of information users and information needs.

6.0 TUTOR-MARKED ASSIGNMENT

Write an essay on the functions and benefits of information centres. Your write-up should be between four and six pages of A4, typed double spaced with 12 points Times Roman.

7.0 REFERENCE/FURTHER READING

Lancaster, F. W. (1979). *Information Retrieval Systems: Characteristics, Testing and Evaluation*. New York: Willey.

UNIT 3 INFORMATION USERS AND USER NEEDS

CONTENTS

1.0 INTRODUCTION

In the last two units you learnt the concepts of information retrieval and the role of information centres in information retrieval. Now you are going to learn something about the user, the characteristics of users that influence their information needs, as well as types of information need.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- define who a user of an information system is
- define a user community
- explain the factors that discourage prospective users from using an information system
- explain the characteristics of users that influence their information needs
- classify information need
- differentiate between "information need" and "demand for information".

3.0 MAIN CONTENT

3.1 Information User Community

Every information retrieval system has its user community. As a matter of fact, an information system is developed with a target audience in mind. The nature of the information system or of the information centre (that constitutes its organisational context) invariably determines the user community. For instance, research libraries (that is libraries serving research institutes) have relatively small homogenous user communities of researchers. A library in an educational institution has for its user community students and teachers. Public libraries have the largest user community and the widest range of interest.

The user community is normally defined by geographic area; institutional affiliation; subject interest; the nature, subject and scope of the information system; and possibly a combination of these.

SELF-ASSESSMENT EXERCISE

What is a user community of an information system?

3.2 The Information User

The information user is, strictly speaking, a person who makes use of an information product, information system or information service. However, in a more general sense, the term also applies to anyone among the people for whom the product, system or service was developed or put in place. This is irrespective of how much the individual is making use of it. Certainly some individuals in a user community will be making intensive use of the service. Some will not be making much use of it, while some will not seem to be using the service at all. People in this last category are usually referred to as non-users, but such classification can be faulted. First of all, a person who has not been using the service being provided may show up any day to request for something. Secondly, there are a host of reasons why a person may not be using a service. For instance, the services may not meet his need, either because it does not match the nature or level of his need or perhaps it is not offered in the right format. Sometimes, the performance of the system may be so poor that a good number of the people who should be using the system look for better alternative sources of information elsewhere. Could such people be regarded as non-users? Certainly, they have not decided that they will not use the system. The problem is with the system.

SELF-ASSESSMENT EXERCISE

Name the three categories of people mentioned above that community of an information system.

3.3 Factors Affecting the Use of an Information System

We want to take a closer look at the factors that affect the use of an information system by those that it was intended to serve. We do not intend to be exhaustive, but to concentrate mainly on factors relating to the system.

Size of population

An obvious factor to consider is the size of the user population. One problem in planning for an information system is to be quite clear on the nature and size of the user population. If other factors are right, a large user population will make a heavy or at least substantial demand on the system. The characteristics of the individuals in the population are also very important. This is an issue that should be treated at length; and so, we shall come back to it later.

Accessibility

Another consideration is accessibility of the system. Physical accessibility has to do with the possibility and ease of the user getting to the location of the system, of gaining entry, to the system, or of having the services of the system delivered to him or her. If for instance, the hours of opening do not suit some people or are not convenient, then the system is not very accessible to them; and they will not make much use of it, if they use it at all. Every kind of bottleneck has to be removed to increase accessibility.

A different kind of accessibility is intellectual accessibility. The level of services and content of information sources must match the intellectual level of users. For instance, in a university library, the undergraduate students use mainly textbooks while post-graduate students and lecturers use journals intensively. This is a reflection of different intellectual levels. An information system that has a user community comprising segments of different intellectual level must be conscious of this intellectual segmentation and ensure that it provides materials to meet the intellectual levels of all its users, otherwise some user category may be "locked out" through intellectual inaccessibility.

Closely related to intellectual accessibility is psychological accessibility. There may be situations that alienate some members of the user community, especially when they feel inadequate or even unwanted. Some of the operators of the system may not be courteous or patient.

Perhaps there is a problem with technology that is complex and which is not supported by adequate user assistance. These are capable of creating psychological barriers.

Cost of using the system

When cost of using an information system becomes higher than what the users are comfortable with, patronage will drop. There are various costs that come into the picture, for instance, cost of making access to the system, cost of subscription, service fees, and in the case of online services, telecommunication charges.

Ease of use

A system that demands considerable time from users will not enjoy much use. The duration of the learning period is equally important. Many people are not prepared to go through a long learning period before they begin to reap the benefits of an information system. They will also not be patient with routines that are complex.

Previous experience

Very often, people's actions are guided by past experience. Anyone who has had an unpleasant experience with an information system would need a lot of convincing to try again. A repeated unpleasant experience can hardly ever be erased.

Value of the services

How useful are the services to the user community? The value the users attach to an information service depends on a number of considerations, including:

- accuracy of information provided
- relevance of the information
- adequacy of the information
- suitability in terms of intellectual level
- suitability in terms of packaging and format
- timeliness
- speed of delivery.

Users’ expectations

An information system has to build up considerable reputation before it can inspire confidence in its users. If the people who are supposed to be using a system do not think highly of it, they will not use it.

3.4 User Characteristics

Much about a person's need of information and his use of information depend on a host of factors that are located in the individual. Age, for instance, could be an important factor. Surely people of different age groups are not likely to need the same kind of information, as their interests will likely be different. Certainly, age will be related to education to a large extent. Even if two persons with different educational background require the same kind of information, the way it is presented might be suitable to one and unsuitable to the other. Inadequate educational level is usually the cause of intellectual inaccessibility of people in respect of specific information systems.

A person's job can be an important determinant of his information need and what information services he chooses to use. The administrative staff in an establishment does not need the same kind of information as the technical staff. It is well known that when a person changes from a job, for instance, from a position as a scientist in the industry to a position as a researcher in a university, his information need changes. Even in the same job, a person's information need and use of information are dynamic and very often depend on the project the person is currently working on and the stage where he is in the project.

A person's official rank has an influence in his information needs and information seeking behaviour. For instance, the information needs of the three management levels, namely strategic (or top) management level, tactical (or middle) management level, and operational level, have different information needs. The top management level requires brief summaries of internally generated data, with, graphical projections as well as brief reports about the external environment of the organisation. The middle management staff require detailed reports of both the internal and external environments to be: able to prepare briefs for top management. The operational staff depend more on procedural information in their job situation.

Subject background is nearly always a factor in the use of subject-oriented information system. Whenever an information system has a subject bias, the subject background of prospective users is defined. The growth rate of the literature is not the same for all disciplines. In the subject areas in which: the literature is growing more rapidly, (usually in science related fields), there is greater pressure on users to keep pace with developments in the field. That pressure will be manifested in the nature and intensity of information needs.

The influence of a person's profession has also been observed. Besides the disciplinary leanings of every profession, there are attitudinal characteristics of the profession that predispose the user td particular sources of information. The disciplinary orientation will necessarily introduce the same kinds of considerations that were covered in subject background.

Other factors include a person's value system, his ambition, and his status in the community,, which could influence his needs of information and attitude to information services, either positively or negatively.

SELF-ASSESSMENT EXERCISE

Name eight factors that determine a person's need of information and his use of information services.

3.5 Types of Information Needs

There are several ways to categorise the information needs of people. One obvious way is to consider what the information will be used for; and so, we can recognise factual information, research information, statistical information, directional information, procedural information, historical information, and so forth.

The types of query that people bring to a document retrieval system provide another way of classifying information needs. While some users request for particular documents by author or title (known item request), others request for documents that can provide answers to the questions they are contending with or that address their areas of interest (subject request). Thus, we have two types of needs here: a known-item need and a subject need.

Lancaster (1979) classified subject need into:

- the need for information to aid in the solution of a particular problem or facilitate the making of a particular decision; and
- information on new developments in a particular field of specialisation.

The first of the two is a current awareness need, which is best satisfied by current awareness services. The need does not have to be strictly defined. Besides, the user is more passive in respect of his role in meeting the need.

The second need is one of retrospective bibliographic search. The user makes his request in well-defined terms because he has defined his problem and he knows exactly what kind of information he needs.

His request can take one of the three forms that express three needs:

1. An answer to a question or factual data. The answer or factual data will normally be supplied from a document, but the user does not have to be given any document.
2. There may be a need for a few relevant documents that address the user's area of interest. This is the most frequent need that users bring to the library.
3 There is also the need for comprehensive literature search, which in most cases is taken to bibliographic databases in the subject area. The people who need this kind of service are mostly researchers.

SELF-ASSESSMENT EXERCISE

In this section you learnt two ways of classifying information needs namely (1) according to what the information will be used for and (2) according to types of query people bring to an information or document retrieval system. With the aid of a sketch diagram, bring out the details of the classification based on the latter.

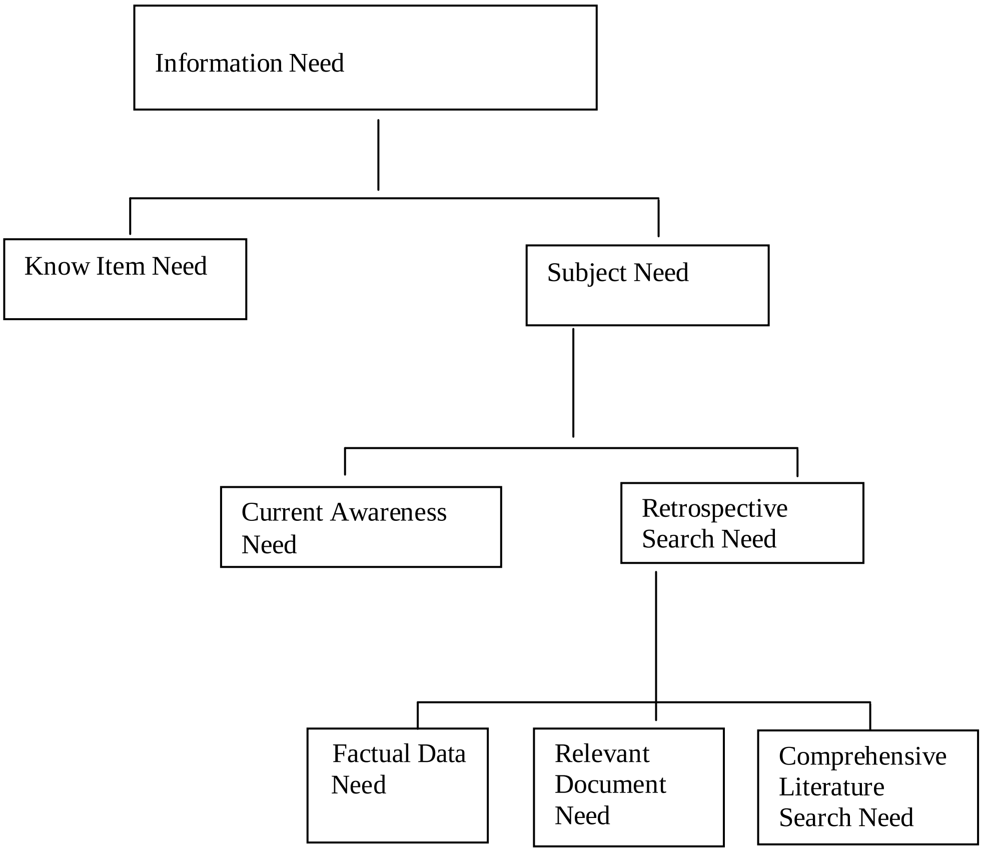


Fig. 1: Types of Information Needs

3.6 Needs and Demands

One big problem in information service is to be able to determine accurately the need of a user community. The problem arises from the fact that the demands being made by users are often wrongly assumed to reflect their needs. The problem has proved to be difficult because:

1. The need of the members of the user community who are not using the service is often not known.
2. It has been found that those who are actually using the service, or making a demand on the service, demand less than they need. In most cases their demand is limited by their expectation of the system's capabilities; that is, what they think the system can deliver.
3. In many instances the needs of a user community are not expressed, simply because the user community is not aware of or does not recognise such needs. Even if some individual do, they may not be motivated to express them in a form of demand. There is no way a system can respond to such latent (unexpressed) needs. So, it is a challenge for an information centre to determine:

- the difference between needs and demands of its user community in quantitative terms
- what types of needs are not converted into demands
- what factors determine whether or not a need is converted into a demand
- how will the demands of users accurately reflect their real information needs.

SELF-ASSESSMENT EXERCISE

Give three reasons information systems do not usually meet all the needs of their user communities.

3.7 User-System Interaction

User-system interaction is the communication between a user and the staff of an information centre through which the user expresses his information need. In an online system that a user logs into without assistance, the user-system interaction is the search strategy formulation process through which the user conveys his need. The user-system interaction is an important issue in information retrieval because of a number of pertinent questions that arise.

1. How accurately does the user express his need? If he is not very clear, his request may be translated wrongly.
2. Does he express his need in more general terms than his need? If he does, a search for him will generate many items that are outside the scope of his interest. That is, a substantial part of what will be retrieved for his request will be irrelevant to his actual need.
3. Does he put his request in more specific terms than his actual need? If he does, a good number of items that are of interest and relevance to him will not be retrieved.
4. Does he express all his need? A need not expressed will not be responded to.
5. Does the vocabulary of the system adequately represent the concepts occurring in the request? If the concepts in the request cannot be easily translated into the index terms of the system, then there is the possibility that the request will be mapped into index terms that do not exactly convey the user's need. The output of the retrieval will be very disappointing.
6. Is the search strategy an accurate representation of the request? If it is not, the output of the retrieval will not be satisfactory.

In an online search, the problem is more on the ability of the user to articulate his needs and to convert his needs into accurate search strategies.

SELF-ASSESSMENT EXERCISE

State four conditions for a user-system interaction to produce the best result.

4.0 CONCLUSION

In this unit, you have learnt the characteristics of users and user community that influence the need for information and the demand they make on an information system, you have also learn type of information need, and the conditions that facilitate successful user-system interaction. You can now appreciate the difficulty of meeting all the needs of the user community of an information system.

5.0 SUMMARY

You have learnt in this unit the factors that constitute major concerns in understanding the needs of users and in maximising the effectiveness of the services of an information retrieval. This has prepared you to learn the principles and techniques of information searching, which will come up in the next unit.

6.0 TUTOR-MARKED ASSIGNMENT

Describe the factors that influence the attitude of members of a target community to the use of an information retrieval system. Your answer should be between four and six pages of typed A4 double-, spaced, 12 points Times Romans.

7.0 REFERENCE/FURTHER READING

Lancaster, F. W. (1979). *Information Retrieval Systems: Characteristics, Testing and Evaluation*. New York: Willey.

UNIT 4 SEARCHING FOR INFORMATION

CONTENTS

1.0 INTRODUCTION

In the last unit, you learnt the characteristics of users and user community that influence their information needs, types of information need, and some fundamentals of user-system interaction. In this unit, you will learn how to retrieve information in different settings.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain informal source of information
- find an retrieve information in the library
- retrieve information from a database
- formulate good search strategies.

3.0 MAIN CONTENT

3.1 Using Informal Sources

The process of searching for information begins with recognising a need. You will recall that we examined the different types of information need in the previous unit. Having recognised an information need, a person should be able to, assess every dimension of the need and be sure he can define and express the need precisely. If, for instance, someone has some doubts or hazy ideas about the history of the amalgamation of the northern and southern protectorates to form Nigeria as it is today, how does he define his need? Is he in need of a kind of revision of the history of the amalgamation or is he feeling at a loss because he has forgotten a specific date? The need for a specific date is different from the need to read up the history of the amalgamation.

When people realise that they need information, where do they go to get it? A lot of people would talk to someone who might be able to supply the information. A scholar would most probably talk to another scholar. A student would ask another student or his lecturer for assistance. A person in a rural community area would ask an elder. This is one way of getting information and it is called informal channel. It is important because the greater part of human communication in everyday life is through the informal, channels. Informal communication does not involve processing of records or manipulation of files. It is prompt and there is immediate feedback. Information retrieval is from a person's mind or memory. There may be a referral process, when somebody who is approached for information refers the person who needs the formation to some other likely source. Our major concern here is about searching for documented information.

3.2 Finding Information in the Library

There are several approaches to finding information in a library. You may use the author-title catalogue, subject catalogue, and possibly a computer-based catalogue.

Using author-title catalogue

Here we shall go through several approaches that you can use to locate information materials in the library. The approach you would use in any given instance depends on whether you want a document, whose author or title you know (known item need) or whether you want whatever documents might be relevant to your need (subject need).

In the first case you would go to the catalogue cabinets labelled "Author-Title Catalogue". You are there now. The author of the book you want is "Iberun O.J." and the title is "Introduction to Psychology". Read the labels on the catalogue trays. The label on each tray indicates the letters or names of the first and last entries in that tray, that is, the first and the last cards in that tray. The first tray starts with A. So you have no problem locating the tray that should contain "Iberun". Starting from the first card, flip over the cards - until you find Iberun, O. J. or until you are sure there is no such name in the catalogue. If there is no such name, it means that the book you want is not in the library. You may want to make sure, so now you are going to search by title. All you should do is to locate the tray where you can, find "Introduction." It will also be one of the "I" trays. You will find "Introduction to Anatomy", then, introduction to various subjects. A short cut is to jump ahead and see whether you have passed the right card or not. Then you move ahead again or move backwards. It is possible you find "Introduction to Psychology" but the author is not Iberun O.J. If that happens, you have to take a decision whether you want to leave or see this other book.

If you want to see this book, then copy down the class mark. It is a code number that is boldly written on the card. Suppose you found "Iberun, O. J. Introduction to Psychology". You would also copy the class mark and, then you would be ready to go for the book. Now, you have the class mark of the book sand so you can locate it easily. If you do not know your way in this library, ask a library staff at the circulation desk and he or she will direct you where to go. The notation (that is, the code) allows the, ordering of materials on the shelves. You saw examples of notations when you studied classification. Just continue scanning the notations on the spines of the books until you see those that look like what you copied from the catalogue. It is likely that many books will have notations that are very close to it. Such books are related. Bring out the books and see the authors and titles. You will eventually find the one you want. If it is not there, go and report at the circulation desk to know where else the book could be. It is possible that the book has been borrowed, or it is awaiting shelving, or else it has been placed on reserve.

Using subject catalogue

Subject search is a little more complex than author-title search. When you are searching by author or title your need is already well defined. You are very sure what you want and no library staff will misunderstand you either. However, when you have a problem to solve or subject matter to explore, then the problem of proper definition comes in. Let us assuming for instance you have an assignment on "The psychological effects of peer pressure on young people." The broad subject is psychology, but you are not interested in the whole of psychology. There are three concepts that together define your need, namely psychological effects, peer pressure, and young people. All three (rather than one or two) together define your need.

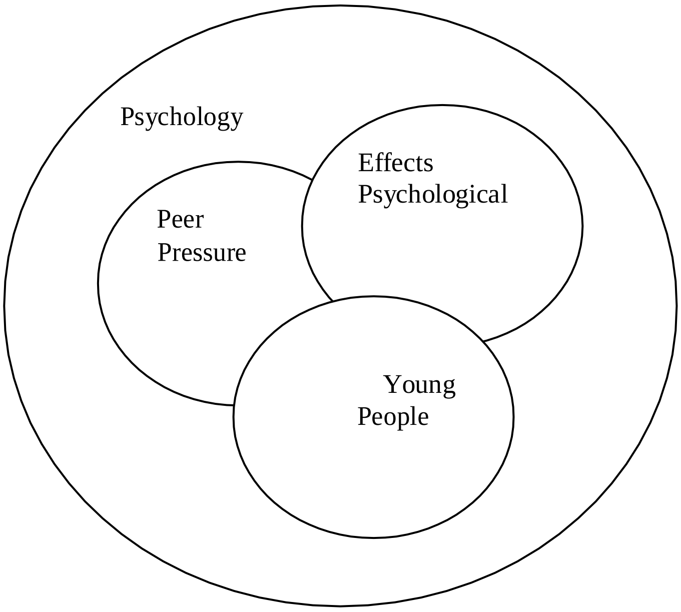


Fig. 1: Relationship of Concepts

With these three concepts, go to the subject catalogue. The trays and cards are arranged alphabetically. Search for each concept. The documents that address all three concepts are the most relevant. Write their class marks. Those that address two out of the three are the next choice that you may consider. The ones that address only one of the concepts are far away from your interest, but still you may want to examine some of them. With the class marks you have collected, go to the shelves to fish out the materials.

Using computer-based catalogue

Using computerised catalogue is the most convenient. A computer-based library catalogue that library users can use interactively is called Online Public Access Catalogue (OPAC). The screen is usually set up to make it easy for the user to type in the terms with which searching will be done, whether by author, title or subject. If you have any difficulty, the library staff in charge of the OPAC will assist you. The results of your searches will be displayed on the screen. Copy the class marks, and then, go to the shelves to locate the materials. If there are no documents to match your search terms, the computer will display "Not found".

3.3 Format of Information Materials

It must be emphasized that the format of the information materials that will be retrieved are not always in, book form. It may be in the form of microform, audio-tape or audio-cassette, audio-visual tape or cassette, maps, newspapers, and all kinds of publications. The class mark will serve as a guide to both the format and the location of the information materials.

3.4 Formation in Databases

The bibliographic databases of most libraries in developing countries are inadequate to satisfy the need for comprehensive search of the literature of the field of interest of information users. The retrieval outputs from comprehensive literature searches, many libraries have to subscribe to external databases.

The selection of databases is an important consideration. The library decides which databases to subscribe to according to the needs of its users but also taking into consideration:

- availability of database
- scope and subject coverage
- cost of using database
- facilities for using database
- extent of overlap with other databases.

Up to 1980 many university libraries in Nigeria were subscribing to major databases, which were delivered in print format. Then the problem of inadequate funding began. Libraries were adversely affected and subscription to such databases stopped. Cheaper alternatives to print version of the databases, namely CD-ROM editions, have since become widely available. In Nigeria, online connection to external databases is still quite expensive and with inadequate infrastructure, unreliable. Most of the libraries that are now subscribing to these databases are subscribing to the CD-ROM versions.

3.5 Query Formation

The process of constructing the syntax used in interrogating a database system is called the query formulation. We shall take a number of examples to show various options in stating one's request for retrieval. Let us return to the concept presented in Fig. 1.

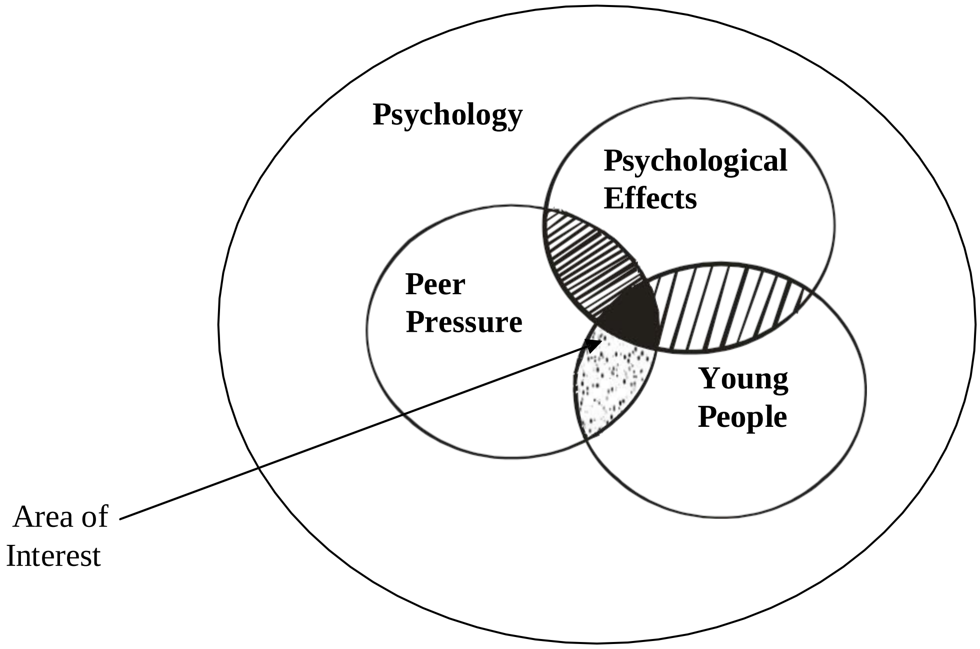


Fig. 2: Query Formulation

A search for information to support the assignment on "Psychological effects of peer pressure on young people" would take into consideration the distinct concepts involved in that need, namely:

- peer pressure
- psychological effects
- young people.

It would also consider the relationship between the concepts. The formulation used will determine the retrieval output. The best formulation is that which insists that all three concepts must be addressed by a document to qualify for retrieval. We can state that request in three different ways.

Approach I

We can go through four steps:

S1	Peer pressure	= 237
S2	Psychological effects	= 728
S3	Young people	= 185
S4	S1 AND S2 AND S3	= 23

(S4 is the intersection between A, B AND C.)

In step 1, we requested for all documents that are indexed with "Peer Pressure" and the number of hits (documents that meet the condition) was 237. In step 2, we request for all the documents that are indexed with "Psychological effects" and the number of hits was 728. In step 3, we requested for all the documents that are indeed with "Young people" and the number of documents found was 195. Then in step 4, we asked the system to reconcile the outputs of the first three steps, picking out only the document that have been indexed with all three concepts, that is ,documents that address all three concepts. Now only 23 documents are retrieved.

Approach 2

S1	Peer pressure	= 237
S2	Psychological effects	= 728
S3	S1 AND S2	= 86
S4	Young people	= 185
S5	S3 AND S4	= 23

S3 is the intersection between A and B, while S5 is the intersection between A, B and C. The problem is that the steps here are too many, but this approach shows the progression in the process of retrieving and filtering records.

Approach 3

We may want to specify all the concepts and conditions in one statement.

S l Peer pressure and Psychological effects AND Young people = 23.

If the system relays its procedure on the screen, you will find that it is, in fact, following Approach 1. As a, mater of fact, unless otherwise instructed it will actually go through more steps than are in Approach 1, especially in a natural language system

S1	Peer	= 942
S2	Pressure	= 1,722
S3	S1 AND S2	= 237
S4	Psychological	= 1,908
S5	Effects	= 2,376
S6	S4 AND S5	= 728
S7	Young	= 394
S8	People	= 449
S9	S7 AND S8	= 185
S10	S3 AND S6 AND S9	= 23

We can improve the output by coordinating some of the words to communicate to the system the meaning we have in mind. Note for instance that documents that are indexed with the terms "Peer" and "Pressure" may not actually be addressing the issue of "Peer Pressure". If we want the system to take each of the three concepts as one term, then we should restate them in Approach 3 as

Peer **w** pressure and Psychological **w** effects and Young **w** people.

S1	Peer **w** pressure	= 148
S2	Psychological **w** effects	= 292
S3	Young **w** people	= 93
S4	S 1 AND S2 AND S3	= 9 .

Besides "AND" the other two Boolean operators are "OR" and "NOT". The Operator OR is used to include alternatives terms (synonyms, near-synonyms, and related terms) that will retrieve equally useful documents. For instance, we may expect that some documents will be indexed with the term "Youths". Then we should restate Approach 3 as

Peer **w** pressure AND Psychological **w** effects AND (Young **w** people OR Youths).

The documents to be retrieved under both Young people and Youths will be more than 185, and the final output is likely going to be more than 9 hits. The operator NOT is used to exclude documents with the stated parameter, for instance

Psychological **w**, effects NOT schizophrenia. This means to retrieve all documents indexed with the term "psychological effects" that are not also indexed with the term "schizophrenia".

What is the use of the three Boolean operators mentioned in this section?

AND Use to indicate that only documents indexed with all the terms joined with AND should be retrieved.

OR Used to include alternative terms that will retrieve equally useful documents

NOT Used to exclude documents indexed with the specified term.

4.0 CONCLUSION

In this unit, you have learnt some of the informal information sources that people resort to in the first instance, and how to find information in the library and in a database. You should now be able to, formulate search strategies to suit your need.

5.0 SUMMARY

Now that you have learnt the details of query formulation, you are ready to learn the relationship between documents and search terms, which will be presented in the next unit in a document-term matrix.

6.0 TUTOR-MARKED ASSIGNMENT

Visit a library and borrow a book on any topic of interest. Write a report on your experience. Your answer should be between three and four pages of A4 typed double-spaced, with 12 points Times Romans.

7.0 REFERENCE/FURTHER READING

Lancaster, F. W. (1979). *Information Retrieval Systems: Characteristics, Testing and Evaluation*. New York: Willey.

UNIT 5 DOCUMENT–TERM MATRIX

CONTENTS

1.0 INTRODUCTION

In the last unit, you learned how to find information in the library and in a database. You should now be able to formulate search strategies to suit your need. In this unit you will explore the relationship between documents and index or search terms as will be presented in a form of matrix.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- construct a document-term matrix
- identify the documents that constitute the membership of a subject class in a matrix
- identify the classes to which a document belongs in a matrix, and
- explain the relevance of the document-term matrix to the performance of a retrieval system.

3.0 MAIN CONTENT

3.1 Concept of Document-Term Matrix

The concept of document-term matrix is well elaborated by Lancaster [1979]. A database links the identifications of documents to index terms assigned to the documents to represent their subject matter. So we can actually regard a database as a document-term matrix. You will recall that in the indexing process, a number of terms are selected as tags for each document. These tags are indications of subject classes to which the document belongs. They also serve as access points to the document.

You will also recall that a database is made up of records [or files of records]

	1	2	3	4	5	6	7	8	9	10
A		x	x			x				
B							x			
C	x				x		x			x
D			x	x						
E	x	x				x				
F	x								x	
G			x					x		
H	x				x			x		
I				x						x
J					x		x			

Fig. l: The Document-Term Matrix

In order to visualize this relationship between documents and terms, let us assume that the entire index terms of a system (that is, the vocabulary) are numbered A, B, C, ... Do not worry about running out of letters. When we get to Z we can continue with AA, BB, CC, ... We can also imagine the records that identify the documents (that is, give the bibliographic details) as numbered l, 2, 3, ... Figure 1 shows a subset of a database. The rows represent the index terms numbered A, B, C, ... and the columns represent documents numbered 1, 2, 3,... So, here we have a matrix. The assignment of any given term to a document is indicated with an 'x' in the appropriate cell. For instance in this subset the first document, that is document 1, has been given the index terms C, E, H.

SELF-ASSESSMENT EXERCISE

Now what are the index terms of the other documents?

The index terms of the remaining 9 documents are:

Document 2: A, E, F
Document 3: A, D, G
Document 4: D, I
Document 5: C H, J
Document 6: A, E,
Document 7: B, C, J
Document 8: G, H
Document 9: F
Document 10: C, I

What we are really saying is that document 1 belongs to class C, class E, and class H. Similarly, we are saying that documents 2 to 10 belong to the classes indicated. Conversely, we can say that since the index term A has been assigned to documents 2, 3, and 6 these documents belong to class A.

SELF-ASSESSMENT EXERCISE

What are the documents in classes B to J?

The classes B to J contain the following documents:

Class B:	7
Class C:	1,5,7,10
Class D:	3,4
Class E	1,2,6
Class F:	2,9
Class G:	3,8
Class H:	1,5,8
Class I:	4,10
Class J:	5,7

This matrix suggests two ways of organising an index. One way is to list all the classes that belong to it (term entry). The other way is to list all the documents in the collection, indicating for each one the classes to which it has been assigned (item entry).

3.2 Subject Search

"According to Lancaster a subject request, translated into the language of the system to allow it to be searched against the index, can be, referred to as a "Search strategy". A searching operation is then the matching of the search strategy against the document-term matrix in order to identify documents whose index term profiles satisfy the logical requirements of the strategy. Now, let us try several subject requests and see how the matching goes.

Request 1 : Document belonging to class A and also class D.
Request 2 : Documents belonging to class E and also to classes F and J.
Request 3: Documents belonging to class E but do not belong to class C.
The documents that satisfy these requirements are as follows:
Request 1 : 3
Request 2 : None
Request 3 : 2, 6

SELF-ASSESSMENT EXERCISE

Which documents will be retrieved for the following requests?

Request 1 : Documents belonging to class D and also class G
Request 2 : Document that belong to class C and also classes H and J
Request 3 : Documents that belong to class B and also class C but do not belong to class J.
The documents that satisfy these requirements are as follows:
Request 1 : 3
Request 2 : 5
Request 3 : None

The document-term matrix shows two ways of searching the database corresponding to term-on-item index and item-on-term index. In a term-on-item index one has to examine serially all the columns of the matrix (that is, document records) in order to identify columns (document records) whose index terms include the terms specified in the search strategies. In an item-on-term index, one has to go row by now, corresponding to the terms specified in search strategy, and identify the member documents.

3.3 Relevance to Retrieval Performance

The document-term matrix can help us appreciate the implications of decisions in the indexing process for the retrieval performance of a retrieval system. Indexing decisions are based on indexing policy characteristics of the index language. You will recall that two of the major characteristics of an, index language are specificity and exhaustivity. Specificity is about the degree of decomposition of subject categories into smaller units and giving tags to such units. The higher the level of specificity, the more the number of classes that will be created. With a high level of specificity, each class will, have fewer members. Exhaustively has to do with the number of index terms selected to represent a', document. The more terms selected to represent a document, the more classes it belongs to and the more the access points from which it will be retrieved. A document may be very important to some category of users, but if it is not indexed with a term that represents their subject category, the document will not be retrieved in a search to meet their interest profile.

Besides these characteristics of the document-term matrix, namely the specificity and exhaustivity of the index language, the other factors that affect the retrieval performance are the quality of requests and, the quality of the search strategies derived from the requests.

4.0 CONCLUSION

In this unit, you learnt the relationship between documents and terms in a database. By now you should be able to construct a document-term matrix, identify the documents that constitute the members of a subject class in a matrix, identify the classes to which a document belongs in a matrix and explain the relevance of the document-term matrix to the retrieval performance of a system.

5.0 SUMMARY

In this unit you have explored the relationship between documents and index or search terms as presented in a form of matrix. The matrix was meant to illustrate the relationship between documents and terms in a database. In the next unit you will learn how to retrieve information from the Internet.

6.0 TUTOR-MARKED ASSIGNMENT

Using a 10 by 1.0 matrix of your own construction, list the membership of each class, and the classes to which each document belongs.

7.0 REFERENCE/FURTHER READING

Lancaster, F. W. (1979). *Information Retrieval Systems: Characteristics, Testing and Evaluation*. New York: Willey.

MODULE 4

UNIT 1 RETRIEVAL FROM THE INTERNET

CONTENTS

1.0 INTRODUCTION

Having learnt how to search for and retrieve information in the library and in databases, you are now ready to learn how to do the same thing on the Internet.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain what the internet is
- use electronic mail
- identify web site addresses
- describe other important facilities on the Internet.

3.0 MAIN CONTENT

3.1 What is the Internet?

The Internet is fondly called the network of networks. It is a medium of communicate made possible by computers linked in networks through telecommunications. The Internet has no national or regional boundaries. It is global and accessible anywhere in the world.

More and more people are discovering that they really need the Internet as a means of communication with people and as a source of information. The electronic mail (otherwise called email) is helping to keep people in regular contact all over the world and to send message to one another. News groups link distant people in discussion fora, people who are interested in the same topics. The Internet is a medium for accessing vast stores of information in every subject area of interest.

3.2 Electronic Mail

For establishing and maintaining contact with people anywhere in the world, for informal discussion, and for quick responses, electronic mail is the most appropriate Internet service. The email facility is serving for the purposes of communication and transfer of information. You can request someone to send to you a journal article in electronic format by email or to do literature search in a database for you and send you the output by email. It will be sent as “attachment” to the email message. Having an access to email facility opens the way for you to use some other Internet services.

In order to use this service, you must have an account with an Internet Service Provider (ISP) or an institution, a telephone line, and a computer with a modem. You may wish to talk to people who are using the Internet to help you identify an ISP that you can use. Please note that having a personal, account with an ISP costs substantial amount of money. However, it is affordable to those who are determined to have it. Besides the initial cost of installing the Internet services software, you will be paying for the 10 or connect charges as well as telephone bills on a continuous basis. If you cannot afford to have your own account, the alternative is to send and receive an email message at a cyber cafe.

To send or receive an email message is not difficult. Unfortunately, we shall not attempt to give detailed instructions here because the interfaces (or your screens) that you will find in various places will be quite different. If you have your own computer and personal account, you will very quickly learn how to get to the message centre and to the blank window screen for writing a message. If you are using a cyber cafe, you just have to let operators get you started. Note that they have to give you a "slot" in the system, activate a clock to keep track of the time you spend on the system and "lock" you out when your time is over.

Once you are in the screen for writing a message, you have to specify the email address of the recipient, the subject of the message, as well as the name and location of a file you want to send as an "attachment" (if any). Some systems will not allow you to begin typing your message until you have specified these. When you finish typing and editing your message you have to ask the system to send it and it will do so.

To receive a message, you go to the screen that contains a list of the messages that have arrived and inspect it. If the message you are expecting or a new message for you is there, you select, open, and read it. If there is an attachment, you open it through appropriate word processing software to read it.

SELF-ASSESSMENT EXERCISE

Visit a cyber cafe and send an email to somebody. You do not have to say mush, but just ask how the person is doing. Remember to go with the person's email address.

3.3 Listserv: Mailing List

Listserv brings members of a discussion group together in an electronic conference (e-conference). The members send their contributions to the address of that group from where they are redistributed to all the members on the fist. Besides using the list as a discussion forum, a listserv group can be used as a resource of obtaining information and answers to questions. Some lists are in fact used for distributing electronic journals and electronic newsletters. There are so many groups spanning almost every area of interest that you should be able to find a list that suits your interest. In order to subscribe to a list, you must know the name of the list, the address of the listserv that manages the list, and how to subscribe.

3.4 Usenet: Newsgroups

Usenet offers what are called Bulletin Board Services of the Internet. It brings together people interested in various topics in what are called newsgroups. Newsgroups are more informal than listserv discussion groins. They constitute a forum for free expression, for posting announcements, and for receiving answers to questions.

3.5 Gopher

Gopher is a powerful Internet facility that makes it possible for people to locate and access a wide range of information resources on the Internet. The interface is a simple menu-driven type, and anyone on the Internet can access and retrieve information located anywhere in the world and which is accessible through the Internet. Information can be retrieved in text files, graphics, or sound. A user can navigate from his computer (root menu) to all the menus and information accessible through gopher (gopher space) in a tree-like manner. In order to use Gopher you must have a gopher client program, which will be communicating with the Gopher server program. Since you will probably be working in a Windows environment you will need a Gopher client program for Windows. A form of catalogue is necessary to help find out what information is available and where. Veronica provides this aid for Gopher. By searching Veronica you can find Gopher files on the Internet.

3.6 Telnet

Telnet allows a user to connect directly to remote computer and access the resources there without, going through layers of menus. With a telnet program, a user can run programs and search databases that are on computers all over the world. The drawback here is that you must connect to a specific host computer using its unique name called domain name. Nevertheless, this should not scare you since you are now familiar with email addresses. A telnet address is equivalent to an email address without the username. To use the telnet is quite easy. As you are most likely to be in a Windows environment, you will go through a menu. Just select and open the telnet program. Supply the host computer name.

An important aid to the use of telnet is a large database of telnet addresses. The program that gives access to the database is called hytelnet.

3.7 File Transfer Protocol (FTP)

The file transfer protocol makes it possible to transfer files between computers on the Internet. By using FTP you can connect to another computer and bring a file from it down to your local host computer. You can transfer just about any kind of file but you must know the names and locations of whatever files you want to get. Big files like databases, graphics and software are usually transferred with FTP.

In order to use FTP you must have the client program in your computer. It is this that will request the server program for FTP service. The FTP directory is organised in a hierarchical file structure. It starts with the root directory that branches into subdirectories. The subdirectories are divided into smaller units. With the huge number of FTP sites, a form of catalogue is necessary to help find out what information is available and where. Archie (for Archives) provides this aid. By searching Archie, you can find files on the Internet.

3.8 Wide Area Information System (WAIS)

The wide area information system (WAIS) is an Internet resource specifically for indexing and accessing the vast stores of information on the Internet. WAIS ties up everything together, including text, electronic books, freelance articles, image and graphics, sound, email addresses, email messages, contributions to a discussion, databases and so forth, in indexes that are continuously updated and, stored for the purpose of retrieval. Indexing is done by keyword, that is, the significant words in a, piece of information. Any of the WAIS sites in the world is accessible through a client WAIS program, which links the user to the host server program. WAIS may be accessed in either of two ways. A user may connect directly to a WAIS site using his client WAIS program; or he may access WAIS server through Gopher or the World Wide Web.

3.9 World Wide Web (www)

The World Wide Web (www) or simply the web, is the main information retrieval resource of the Internet. In several ways, the web has something in common with Gopher. They were developed about, the same time (Gopher in the United States and the web in Switzerland). The intention was the same: namely, "to provide a single interface to many different kinds of information and to be able to link them together.” However, the web added a very useful feature called hypertext. A hypertext may be defined as a document that contains references or links to related documents. That means that every piece of information in the web has links to other documents that contain similar information. A user can follow these links and retrieve considerable amount of information.

The web was released in 1991 the same year as Gopher. From 1993, the web started gaining attention worldwide with the development of two browsers (programs providing an interface for accessing the web) namely MOSAIC and Lynx. MOSAIC is a graphical interface (with menu screen and little picture frames representing the options available for selection). MOSAIC was developed at the University of Illinois in the United States, and it became a very popular client program. It actually went beyond hypertext to multimedia capability. Along with the links between related documents, pictures, sounds, and even movies were incorporated. Besides MOSAIC and Lynx, other client programs have since come into the market. For Windows environment, the most popular clients currently are Netscape and Microsoft Internet Explorer. Internet Explorer users are familiar with two popular search engines called Yahoo! and Google.

Like many of the other Internet facilities, the web is based on what is called a client/server architecture or model. The client program is the program that is running in your computer. It takes your requests to the server program, which is running in the computer that provides the web service you want. The client program retrieves the information and displays it on your screen. Searching can be done by typing in either the name of the web site or the key words of the subject of interest.

The name of a web site is called Universal Resource Locator (URL). A URL is made up of method, host computer, and pathname. The method is the kind of protocol used to retrieve the document. For web document in HTML format the method is “http". For Gopher documents and directories it is "gopher" while for FTP connection it is “ftp". Examples of URLs are

http: //www.bellanet.org/reseach
gopher: //gopher. mudge.com/list
ftp: //ftp.univa.edu/projects

The host computer is the computer that stores the information being requested. The specification of the host computer begins after the two forward slashes and ends before the first single forward slash. From the first single slash to the end is the pathname, which is made up of the directory (and perhaps a series of subdirectories) and the files.

Since we mentioned something about HTML format, let us explain it. HTML means Hypertext Mark-up Language. There is the notion of "marking up". What actually happens is that the codes for formatting and linking are imbedded within the text. The text, which contains links to other documents or part of the same document, is usually underlined or highlighted in another colour. One has to click on the underlined or highlighted text to follow the link.

An important question that requires some attention here is "How do I know the addresses of web sites? This is really no problem. These days most organisations that have a web site supply their web site address in their documents. Many Internet guides also carry directories of web sites though they are far from being comprehensive. If you know the web site address you need to search, you are in order. If, however, you do not know any address, you are still in order. In this case you have to search by subject, using keywords or concepts just the same way you search library catalogues or databases with appropriate subject terms when you do not know the authors or titles of particular documents.

4.0 CONCLUSION

This unit has exposed you to the information resources on the Internet. You should now be able to explain what the Internet is, use electronic mail, identify web site addresses, and describe other important facilities on the Internet.

5.0 SUMMARY

With this unit, you have now completed your course materials on information retrieval. In the next two units you will learn how to evaluate information storage and retrieval systems.

6.0 TUTOR-MARKED ASSIGNMENT

Find out the web site addresses of ten Nigerian dailies and magazines.

7.0 REFERENCE/FURTHER READING

Keiko, Pitter *et al.* (1995). *Every Student's Guide to the Internet*. New York: Mcgraw-Hill.

UNIT 2 EVALUATION OF INFORMATION SYSTEMS AND SERVICES: PART I

CONTENTS

1.0 INTRODUCTION

In the previous units, you learnt the principles and practice of information storage and retrieval. In this unit (and in the one that will follow) you will learn how to evaluate information retrieval systems and services.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain the purpose of evaluation
- define evaluation of effectiveness
- enumerate and explain evaluation criteria
- describe the steps in an evaluation programme.

3.0 MAIN CONTENT

3.1 Purpose of Evaluation

Every information system is under continuous evaluation. The evaluation may be formal or informal. In an informal way, the users or those who are supposed to benefit from an information system usually assess the services being provided and decide whether they are worth the cost in money, time or effort demanded. A formal evaluation programme demands some systematic approaches based on accurate measurements. A formal evaluation could be carried out at four levels, namely:

- evaluation of effectiveness
- evaluation of benefits
- evaluation of cost-effectiveness
- cost-benefit evaluation.

3.2 Evaluation of Effectiveness

The aim of evaluating effectiveness is to determine the extent to which a service meets the need of its users. The problem here is that we do not often know the exact needs of the intended users of an information service. What are often known and measured are the demands or expressed needs of a user group. The unexpressed or latent needs of the current users and the needs of the intended users who are hot currently using the service are usually not known and so, are not taken into account.

Evaluation of effectiveness is basically a consideration of user satisfaction. Evaluation criteria include cost, time, quality, availability, and accessibility.

SELF-ASSESSMENT EXERCISE

What is the purpose of effectiveness evaluation? What is the major problem of the usual approaches to the evaluation of effectiveness?

Cost

The first consideration is the financial commitment users have to make to enjoy a service. They may have to pay for subscription, searches, or supply of documents. The issue here is to assess the cost in: relation to the benefits of the service. The cost must be seen to be reasonable from the point of view of users.

A second consideration in cost implication is the effort demanded of users. The service may require that users put effort into learning the use of it. Experience shows that the amount of effort required to learn to use a service could be a critical factor in the amount of use of the service. The effort required to learn the intricacies of the service or interface with the system is one thing. Another is the effort in actually using the system. For instance, how much effort would a user put into a searching section in order to retrieve a fair amount of information? Does the system provide a transparent interface to a number of databases such that the user does not have to try to access each database separately? Users are really having easier time now than previously in searching multiple databases through online or CDROM retrieval systems. The format of an information product may also require some effort to use the product. For instance, a document in microform is far less convenient to use than one in printed form. Whether or not a service will be used may largely depend on the ease of using it.

SELF-ASSESSMENT EXERCISE

List the cost considerations mentioned in this section.

Availability

A user is not just interested in knowing that a library has the information sources that he needs or satisfied with a handsome list of references retrieved from a database. He certainly needs these as a first step. His real desire is to get the documents. Availability is about placing the physical documents at the disposal of the user.

Availability in a library may be a function of coverage or size of collection or otherwise that of circulation factors. Many libraries in developing countries are usually inadequately equipped. They acquire a negligible proportion of the materials in their areas of mandate. So, users very rarely find enough materials to satisfy their needs. The richer libraries often have more users than they can provide for. With the greater level of circulation, materials are very often out on loan, and physically not available.

Computer-based databases have until recently concentrated on providing just references. Some also provided abstracts. An abstract is one more step towards the full information in a document. After reading the abstract, the user may want to have the full document. The level of frustration in not being able to get the physical documents indicated in a bibliographic retrieval is very high for information users in developing countries. It is interesting to know that full texts are being included in more and more databases, most of which are available in CD-ROM.

SELF-ASSESSMENT EXERCISE

What are the constraints to availability of documents especially in developing countries?

Accessibility

Accessibility as an evaluation criterion takes the concern for user's satisfaction beyond availability. The question here is, "How easy is it for a user to gain entry to a system or the location of the information products or documents?” Accessibility is often reduced by security measures or unsuitable hours of operation, or inconvenient location.

SELF-ASSESSMENT EXERCISE

Discuss the factors that create accessibility problem.

Time

When considering time factors, the question is, "How long does the requester for information have to wait?” If, for instance, the request is for bibliographic retrieval, how long will he have to wait to receive the references? This is a pertinent question in a situation in which searches are made for users, and usually in batches.

Waiting time still comes up with online services when users have to queue up to use the system. At the present stage of, development of online connections in Nigeria, especially Internet services, queuing to use the services is a familiar phenomenon.

Requesters for documents also have to wait for varying lengths of time to receive the documents. Instances of long waiting periods are usually associated with retrieving documents that are not directly accessible to users. The staff on duty may be over-stretched and may have to collect requests in batches or otherwise fetch documents in batches. Much longer waiting periods are associated with services in which materials have to be recalled from one user in order to meet the need of another.

SELF-ASSESSMENT EXERCISE

What are the time factors that have been addressed in this section?

It must be realised that the amount of waiting time that may be considered tolerable may vary from one situation to another. A person who is waiting for a document that has been borrowed may be able to tolerate one week or more, while somebody who comes to a library to use materials that are not on open access may be very impatient after waiting for thirty minutes to receive the materials. Similarly, a user who needs comprehensive bibliographic search may be willing to wait for over two weeks while, someone who needs just a few relevant references and the documents to continue the work he has on hand may find two or three days of waiting intolerable.

Quality

The main quality considerations that we need to address are the coverage of the system, recall (completeness of output), precision (relevance of output), and indexing factors.

Coverage

Coverage is usually readily considered in evaluating libraries and databases. The collections in a library may be evaluated against certain standards for such libraries or against growth in publication. Coverage implies both subject areas included and intensity of acquisition in the subject areas.

These considerations are also relevant in the evaluation of databases. In addition, the following considerations are quite important.

- Number of sources: This is usually the number of books, journals and other materials that have been scanned to produce the bibliographic records.
- Type of sources: Usually the format of materials covered in a database is specified. They include books, journals, reports, ephemerals and other grey literature, and so forth.
- Number of items: This is the total number of references or bibliographic records in the database.
- Time span: The dates covered by a database. If for instance the time span of a database is 1975 to present, it means that it covers materials published from 1975 to the present day.
- Completeness in relation to user needs. This is difficult to determine. However, it should be obvious from th6 level of satisfaction reported by users if the coverage is good enough.
- Uniqueness and overlap: Uniqueness is the proportion of a database that is not covered by other databases. Overlap is the proportion of a database that is also covered by other databases.

SELF-ASSESSMENT EXERCISE

We have said that coverage is an important issue in evaluating library collections and databases. List the factors involved in the evaluation of coverage.

Recall and precision

The term "recall" refers to a measure of whether or not a particular item is retrieved or the extent to which the retrieval of wanted items occurs (Lancaster 1979). If a user wants a particular document, recall is simply a question of whether the document is retrieved or not. When a request is for a comprehensive search, then the consideration is the extent to which all the relevant documents, or references to them, are retrieved. This measurement is usually expressed as "recall ratio". A recall ratio of 90% means that 90% of all the relevant items were retrieved.

The term "precision" refers to the extent to which the items retrieved are actually relevant. We found earlier that various indexing factors could create situations in which documents that are of marginal interest to a need or that are completely irrelevant may be retrieved. When evaluating retrieval performance of a system, the question then is, "What proportion of all the items that are retrieved are really relevant to the need intended to be met?" Let us consider a case when a literature search generated 80 items. If the requester finds only 20 out of them relevant, then the precision ratio is 25%.

The relationship between system relevance prediction and user relevance decisions is illustrated by Lancaster in the table below. A matrix of the two creates four quadrants (a, b, c, d). To every search strategy the systems decides what items of the database are relevant and retrieves them (a + b) while holding back the rest (c + d). The user can also see the whole contents of the database as comprising what is relevant to his search (a + c) and what is irrelevant (b + d).

Ideally, the system ought to retrieve all the relevant items namely a + c and exclude b + d. But, the system is not perfect and so it correctly retrieves a and correctly rejects d but erroneously retrieves b and erroneously rejects c. If, indeed, it retrieves a + c, then the recall ratio would be 100%, but since it retrieves only a, then the recall ratio is:

$$\frac{a}{a + c}$$

$$\frac{\text{That is number of relevant document retrieved}}{\text{Total number of relevant documents in the collection}} \quad x \quad 100$$

Ideally b, which is just noise, should be zero. That would mean 100% precisely ratio. However, b is rarely zero and the precision ratio is:

$$\frac{b}{a + b}$$

That is, $$\frac{\text{Number of relevant documents retrieved x}}{\text{Total number of documents retrieved}} \quad 100$$

Table 1: Results of a Literature Search (Lancaster 1999)

System relevance prediction	User relevance decisions	a+b+c Total
Relevant Not Relevant	**Relevant** **Not Relevant** Retrieve a (hits) b (noise) Not Retrieved c (Misses) d Correctly rejected	a + b c + d
	Total a + c b + d	

The precision ratio and recall ratio express the filtering capacity of a retrieval system that is the ability to retrieve what is needed and hold back what is not. It must be noted that recall and precision tend to be related inversely; the higher the recall ratio, the lower the precision ratio and vice versa.

SELF-ASSESSMENT EXERCISE

If a = 36, b = 54, c = 45, d = 1964
How many items were retrieved?
How many items were relevant in the database?
How many items were in the database?
What is the recall ratio?
What is the precision ratio?

Indexing

Indexing factors play a very important role in the performance of a retrieval system, and so they should be investigated in an evaluation project. The following factors should be considered:

a. Degree of Vocabulary Control

Vocabulary control or lack of it has a number of implications for search formulation. If there is strict control, then the user is obliged to use only the terms allowed. The use of natural language gives much more freedom of choice of search terms, but the user must be able to formulate appropriate search strategies using all possible terms, including synonyms and near-synonyms, for a good output.

b. Specificity of Vocabulary

Specificity is an important factor in precision. The greater the level of specificity, the greater is the level of precision. Using broader terms as indexing terms allows more items to be retrieved, but many of the items are not likely to be relevant.

c. Exhaustivity

Exhaustivity is a measure of number of index terms selected for any one document. The more terms that are selected, the greater is the likelihood that a given document will be retrieved. However, it is necessary to ensure that concepts that are only marginally or superficially treated, are not selected, otherwise the quality of output in terms of precision will suffer.

d. Accuracy and Consistency

The integrity of an information retrieval system depends to a large extent on the accuracy and consistencies of the indexing operation. Accuracy is a measure of the extent to which the, indexing is free from indexer errors. Indexer errors are of two types:

- omission of terms that are necessary to describe important aspects of the documents
- use of terms that are inappropriate to the subject of the document.

The implications of these errors for retrieval performance are quite obvious. Omissions normally lead to recall failures. That is, the documents for which important terms are omitted from the indexing are likely to remain unretrieved in a number of searches to which they are actually very relevant. The use of inappropriate or non-specific terms could cause either the retrieval of irrelevant items (precision failures) or make it impossible to retrieve the affected documents (recall failure). Therefore, indexing errors should be investigated.

Consistency is the extent to which two or more indexers agree on the choice of terms needed to represent the subject matter of a particular document (inter-indexer consistency) or the extent to which the same indexer, at different times, agrees with himself on the choice of terms to represent the subject matter of some document (intra-indexer consistency). The consistency achieved by a group in indexing a particular collection of documents is influenced by many factors, including:

- The exhaustivity of the indexing
- The type of vocabulary used
- The size and specificity of the vocabulary
- The experience and training of the indexers
- The subject field of the documents
- The types of indexing and provided (Lancaster 1979).

e. Searching aids

The retrieval performance could be enhanced with suitable searching aids, especially a thesaurus.

SELF-ASSESSMENT EXERCISE

Why are indexing factors important in determining the performance of an information retrieval system?

3.4 Steps in an Evaluation Programme

The major steps involved in an evaluation programme are as follows:

a) defining the scope of the evaluation
b) designing the evaluation programme
c) execution of the programme
d) data analysis and interpretation of results
e) modifying the system or service.

Defining the scope of the evaluation

The definition of the scope of the evaluation programme involves the preparation of the set of questions that the evaluation programme is intended to answer. The definition of scope is precisely a statement of what is to be learnt through the study. It is against the definition of scope that the evaluator will design the evaluation programme.

Designing the evaluation programme

The design of the evaluation programme involves the preparation of a plan of action for the purpose of gathering the data to answer the questions raised in the definition of scope. For each question in the definition of scope, the evaluator must determine how to collect the data needed. Some questions may require that data be collected from the system as it is presently, while some other questions may require some modifications in aspects of the system to make it possible to generate the kind of data needed. So, while some questions dictate a systematic and controlled observation of the system, others will require well-established procedures of experimental design.

Execution of the programme

It is at the stage of execution of the evaluation that data are collected. A pilot form of data collection may be needed to validate the methods for data collection and analysis. Data collection is likely to take considerable amount of time.

Data analysis and interpretation of results

The task in data analysis is to reduce and manipulate the data in a way to answer or contribute to the answering of the questions for which they were collected. When he has finished data analysis, the evaluator must present his results in a report that explains the results and the significance of the findings as well as make recommendations for the improvement of the system.

Modifying the system or service

In the light of the result of the study, some modifications may have to be made on the system or service that was evaluated. At least the recommendations made should be carefully considered and implemented.

4.0 CONCLUSION

In this unit you have been introduced to the evaluation of information retrieval systems and services. Now you should be able to explain the purpose of evaluation, define evaluation of effectiveness, enumerate and explain evaluation criteria, and describe the steps in an evaluation programme.

5.0 SUMMARY

This unit introduced the topic of evaluation and then treated at length the issue of evaluation of effectiveness. The next unit will take you to other areas of evaluation.

6.0 TUTOR-MARKED ASSIGNMENT

Write a report entitled "An Evaluation of the Services of a Library." In order to get your facts, go to a library and observe its services. Your report should include:

- cost of services
- availability of materials
- accessibility to materials
- response time to requests
- quality of services.

You are not to mention the name of the library. Your report should be between four and six A4 pages.

7.0 REFERENCE/FURTHER READING

Lancaster, F. W. (1979). *Information Retrieval Systems: Characteristics, Testing and Evaluation.* New York: Wiley.

UNIT 3 EVALUATION OF INFORMATION SYSTEMS AND SERVICES: PART II

CONTENTS

1.0 INTRODUCTION

In the last unit you were introduced to the purpose of evaluation of information systems and services. This unit will bring our overview of the subject matter to completion.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain the meaning of cost effectiveness
- explain the purpose of cost-effectiveness evaluation
- describe the steps involved in cost-effectiveness analysis
- explain the meaning of benefits
- explain the meaning of cost-benefit evaluation
- distinguish between relevance and pertinence.

3.0 MAIN CONTENT

3.1 Evaluation of Cost-Effectiveness

Cost effectiveness is the relationship between level of performance and the costs involved in achieving it. Cost-effectiveness evaluation, therefore, aims at relating measures of effectiveness to measures of cost. The usual procedure is to, cost all the alternative methods that could be adopted to obtain a desired level of performance, and then determine which method is least expensive.

The determination of costs could be quite complex. Usually the cost of an information service comes at the input stage so the cost can be measured in terms of input of resources. Under cost, we would need to consider both the costs that are relatively fixed, such as purchase or rental of equipment, developmental costs, acquisition of database, indexing costs, as well as costs that are relatively variable. Costs may be variable either because of variable number of transactions or as a result of changes in the mode of operation of the system. Examples of changes in the mode of operation of retrieval system include shifts in emphasis in the modes of interaction with users, modes of interaction with the database, adding or eliminating some output manipulation tasks, and changes in the professional level of the personnel conducting the searches.

We can actually define evaluation of cost effectiveness as a study of the extent to which available resources are so allocated that the maximum possible return is achieved for the investment made. A cost effectiveness analysis is done to determine which is the least expensive of several alternative methods for achieving a particular level of service.

The cost effectiveness of a service can be improved. This can be done either by maintaining the present level of performance but reducing the cost of achieving it, or by keeping costs constant but raising the level of performance. The cost effectiveness would also improve if it were possible to raise the level of performance while reducing costs.

Steps in cost-effectiveness analysis

Hitch and McKean (1960) proposed the following five steps in cost-effectiveness analysis.

a. definition of objectives
b. identification of alternative methods
c. determination of the costs of the alternatives
d. establish cost-effectiveness models
e. establishing a criterion for ranking the alternatives.

The performance objectives that must be met must be clear and precisely stated. There must be several methods of meeting each objective. These must be identified.

An assessment of all the costs associated with each alternative method must be done. Here, we have to have one or more models that relate the costs of each alternative to an assessment of the extent to which each could assist in attaining the objectives. The models may take the form of mathematical expression or simply verbal formulation.

It is necessary to rank the alternatives in order of desirability so that the most promising can be chosen. The criterion should provide a method of weighing estimated costs against estimated effectiveness.

Cost effectiveness and indexing policies

a. The amount of time expended, on the average, in the indexing of a document.
b. The level of exhaustivity adopted in indexing, that is, the number of index terms assigned, on the average, per item.
c. The professional level of personnel used in indexing
d. The need for an indexing revision procedure.

Experience shows that the decision on the most appropriate level of exhausitivity to adopt is perhaps the most difficult problem with respect to indexing policy. The difficult question is, "How many index terms, on the average, should be used?" The more exhaustive the indexing, the greater the recall of the system is likely to be, but then the precision will likely decline. Given any set of documents, index language, and requests, there is an optimum level of exhaustivity of indexing. A cost effectiveness analysis makes it possible to find this optimum level; that is, the point of diminishing returns after which the addition of further terms is largely unproductive.

The time allowed for indexing a document has an implication for both exhaustivity and indexing accuracy. More time is required for a greater level of exhaustivity. The indexer is likely to make more errors if he has to index more documents in a given period of time. As we have already said, such errors will include omission and the assignment of inappropriate terms. A cost-effectiveness analysis relating to indexing time must take into account the effects of time allowance on exhaustivity and accuracy. Another related question is the necessity for an indexing revision process, that is, the review of the work of one indexer by a second person. The necessity of this depends on:

a) the amount of error occurring in the unrevised indexing
b) the amount error that is corrected by the revision operation
c) the estimated effect of indexing error, revised and unrevised, on retrieval performance
d) the cost of revision.

The level of personnel required to undertake the indexing is another aspect of the indexing process that requires cost effectiveness analysis.
It is an obvious fact that the higher the professional level, of the indexer, the higher the indexing cost. How professionally qualified or how senior must the indexer be? This depends on:

The complexity of the subject matter being handled.

a. The type of index language used. Free keyword indexing may require less skilled indexer than the use of a classification schedule or a combination of relational indicators. It is worth noting that more and more journals now require authors to supply keywords for their articles.
b. The exhaustivity and specificity of the indexing. The greater the technical detail indexed, the greater the need for subject expertise.
c. The stage of system development. In a system using controlled vocabulary and at the early stages in the indexing of a collection, virtually every indexing decision that is made is intellectual. However, with time, the authority file that has been created and the entries in vocabulary provide adequate guide such that less skill is required in the indexing process.
d. The quality of tools provided to aid the indexing process.
e. The quality of the indexing training programme.

Cost-effectiveness analysis of the indexing aspect of an information retrieval system could be done by having various sets of documents indexed by personnel of different levels and comparing the resulting indexing with a standard indexing for the test documents.

Trade-off between input and output costs

Cost-effectiveness analysis of an information system involves a study on payoff factors, trade-offs, break-even points, and diminishing returns. Cost-effectiveness analysis may be carried out on aspects of an information system, including database, indexing, index language and searching procedures. We do not intend to describe the cost-effectiveness analysis on all these aspects. It will be enough to describe the considerations in cost-effectiveness analysis on the indexing subsystem.

Almost invariably, economies in input procedures result increased burden in the output procedures and, therefore, increased output costs. Conversely, greater attention to input procedures resulting in increased input costs, normally leads to improved efficiency and reduced output costs. Lancaster enumerated a number of other tradeoffs, including those between a carefully controlled and structured index language and free se of uncontrolled key words.

The controlled vocabulary requires effort in construction and maintenance and is more expensive to apply in indexing. It takes longer, on the whole, to select terms from a controlled vocabulary, which may involve a lockup operation, than it does to assign keywords freely. Moreover, keyword indexing may require less qualified personnel than the use of a more sophisticated controller vocabulary. The controlled vocabulary, however, saves time and effort at the time of output. Natural language or keyword searching, without the benefit of a controlled vocabulary with classification structure puts increased burden on the searcher, who is forced to think of all possible ways his subject interest could be expressed by keywords or natural language terms. In the same way, the uncontrolled use of keywords may lead to reduced average search precision and thus may require additional effort and cost in output searching.

3.2 Evaluation of Benefits

The aim of benefit evaluation is to determine what impact an information service has on its users. The evaluator wants to find out to what extent the users are benefiting from the service. The difficulty that goes with the evaluation of benefit is that benefits cannot easily be reduced to quantitative terms. Evaluation of benefits therefore tends to be subjective unlike evaluation of effectiveness, which is based on objective quantitative measures. It seems reasonable to assume that there is a direct relationship between the effectiveness of a service and its benefits. For instance, it could be assumed that a community of users is deriving more benefits from a service that is 90% successful in meeting the demands of the users than if it were only 40% successful.

SELF-ASSESSMENT EXERCISE

Why is it necessary to carry out an evaluation of benefit of an information service?

3.3 Evaluation of Cost Benefit

Cost benefit is the relationship between the benefits of a particular product or service and the cost of providing it. Benefits are generally more difficult to measure than effectiveness. In a commercial sense, however, benefits are understood to be return on investment; but then, there are many benefits that are not, as tangible as that. A cost-benefit study is aimed at exploring the relationship between the benefits of a service and the cost of providing it. With the difficulty of measuring benefits, cost benefit studies are usually not easy to carry out.

Some possible criteria for establishing a cost-benefit ratio for information services include:

a) Cost savings through the use of the services as compared with the costs of obtaining needed information or documents from other sources.

b) Avoidance of loss of productivity that results if information sources were not readily available.

c) Improved decision making or reduction in the level of personnel required to make decisions.

d) Avoidance of duplication or waste of research and development efforts on projects which either, have been done before or have been proved infeasible by earlier investigators.

e) Simulation of invention or productivity by making available the literature on current developments in a particular field (Lancaster 1979).

With the difficulties involved in measuring benefits in quantitative terms, most of the studies on assessment of benefits have been restricted to asking users their opinions about the benefits of the services provided. A questionnaire on user perception could generate useful data but the results cannot be said to be completely objective.

SELF-ASSESSMENT EXERCISE

What makes cost-benefit analysis difficult to do? How have researchers been assessing benefits of information services?

3.4 Relevance and Pertinence

It must be noted that relevance judgements made on the basis of the relationship between the, documents retrieved and request statements do not really tell us anything about the degree of success achieved in meeting the information needs of users, whether we consider "actual" need or "recognised" needs. There is a difference between document-request relevance judgements made by intermediaries or even a panel of judges and documents-information need value judgements made by the requester himself. Certainly only the requester can decide whether or not a particular document contributes to the satisfaction of his information need. We could use the term "relevance" to refer to a relationship between a document and a request, based on the subjective decision of one or more individuals; and pertinence to refer to a relationship between a document and an information need. The decision, in this case, is made exclusively by the person having the information need.

4.0 CONCLUSION

In this unit, you have learnt a number of evaluation measures and procedures. You should be able to explain the meaning of cost effectiveness and the purpose of cost-effectiveness evaluation, describe the steps involved in cost-effectiveness analysis, explain the meaning of benefits and cost-benefit evaluation, as well as distinguish between relevance and pertinence.

5.0 SUMMARY

This unit has brought to a close our consideration of the elements of evaluation of information systems and services. This unit also brings to completion your introductory course work in information storage and retrieval. You are now ready to proceed to the more advanced course work in this area.

6.0 TUTOR-MARKED ASSIGNMENT

Write a report of between four and six pages on "My Experience with Searching the Internet." For the purpose of this assignment you are to search "Internet for Information Industry in Africa." You are to evaluate your research result as well as the capability of the Internet to meet your information need. Your evaluation should include:

- cost of the service
- cost effectiveness
- availability of information
- adequacy of information
- response time
- quality of service.

7.0 REFERENCES/FURTHER READING

Hitch, C. J. & McKean, R. (1990). *The Economics of Defense in the Nuclear Age*. Cambridge, Mass.: Harvard University Press.

Lancaster, F. W. (1979). *Information Retrieval Systems: Characteristics, Testing and Evaluation*. New York: Wiley.