



**KARNATAKA STATE OPEN UNIVERSITY**  
**MUKTHAGANGOTRI, MYSORE -570 006**

**Master of Library and Information Science**  
**M.Lib.I.Sc - 5**

**Information Systems:  
Architecture and  
Retrieval**

**BLOCK - 1**

**BLOCK**

**1**

---

**INFORMATION RETRIEVAL SYSTEMS**

---

---

**Unit -1**

**COMPONENTS OF AN IRS**

---

**Unit -2**

**SEARCH STATEMENT**

---

**Unit -3**

**DIFFERENT TYPES OF QUERY FORMULATION**

---

**Unit -4**

**SEARCHING AND SEARCH PROCESS.**

## INSTRUCTIONAL DESIGN AND EDITORIAL COMMITTEE

### COURSE DESIGN

**Prof. D. Shivalingaiiah**

**Chairman**

Vice Chancellor  
Karnataka State Open University  
Mukthagangotri, Mysuru-570006

**Prof. M. Mahadevi**

**Convener**

Dean (Academic)  
Karnataka State Open University  
Mukthagangotri, Mysuru-570006

### COURSE COORDINATOR

**Shilpa Rani N R**

Chairperson

Department of Studies in Library and Information Science  
Karnataka State Open University, Mukthagangotri, Mysuru-570006

### COURSE EDITORS

**Prof. M A Gopinath**

Professor (Retd.) in LISc  
DRTC, ISI Building, Mysore Road,  
Bangalore

**Prof. A Y Asuudi**

Professor (Retd.) in LISc  
Bangalore University  
Bangalore

**Dr. N. S Harinarayana**

Senior Lecturer  
Dept. of Library & Information Science  
University of Mysore, Mysore -06

**Prof. V. G. Talwar**

Professor in LISc  
Dept. of Library & Information Science  
University of Mysore, Mysore -06

### COURSE WRITER

**Dr. P G Tadasad**

Lecturer (Sr Scale)  
Dept. of Library & information science  
Gulbarga University, Gulbarga

### BLOCK EDITOR

**Prof. V G Talawar**

Professor  
Dept. of Library & information science  
University of Mysore, Mysore -06

### PUBLISHER

**Registrar**

Karnataka State Open University  
Mukthagangotri, Mysuru-570006

Developed by Academic Section KSOU, Mysore

**Copy Right: KARNATAKA STATE OPEN UNIVERSITY, 2017**

© All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from Karnataka State Open University.

This courseware is printed and published by The Registrar, Karnataka State Open University, Mysuru for limited use only. No individual or collaborative institution can use / print / distribute in any form without the written permission from KSOU. For user rights of this content and for other queries contact The Planning and Development Officer, KSOU, Mysuru 570 006.

Digital delivery of this courseware is also available for those who opt. For more details visit

[www.ksoustudymaterial.com](http://www.ksoustudymaterial.com) or [www.ksoumysore.edu.in/digitalcontent](http://www.ksoumysore.edu.in/digitalcontent)

**M.Lib.I.Sc – 5 : Information Systems: Architecture and Retrieval**  
**Block – 1 : INFORMATION RETRIEVAL SYSTEMS**

---

**Block Introduction**

Bill Gates the computer Wizard, bringing the analogy of online supermarket and home delivery in his address to the 1995 DLF (Digital Library Federation) Business Forum said Users will want their information accessible to them when they want it. Driving to a library which may not be open, hoping to find the information wanted is not the future model. The 24 hours reference library and IT services will supersede the check out circulation clerk.... The future can be utopian cybernetic or Orwellian in vision... we have a myriad of individual net accesses a plurality of Cyber villages but with most users ghettoized in their dedicated information and entertainment channels”. User becomes the focal point and he needs access to information and entertainment at his or her desktop in the office or at home”.

Developments in IT, the Internet and the Web have led to the development of the new era of electronic and digital libraries. Research in digital libraries is taking place in both developed and developing countries on account of rich funding by various national and international bodies. Digital libraries offer unique ways of recording, preserving and propagating culture in multimedia form. Apart from being organizations that preserve traditional culture (language, art, music, folk art etc) digital libraries serve to forward the frontiers of science. As Witten and Bainbridge opine “if information is the currency of the knowledge economy, digital libraries will be the banks where it is invested. Indeed Gothe once said that visiting a library was like entering the presence of great wealth which was silently paying untold dividends”.

**Prof. V G Talawar**

---

## **Unit – 1: COMPONENTS OF AN IRS**

### **Structure**

1.0 Objectives

1.1. Introduction

1.2. Purpose of an IRS

1.3. Functions of an IRS

1.4. Service oriented IRS

1.5. Components of an IRS

1.6. Check your progress

1.7. Summary

1.8. Glossary

1.9. Questions for self study

1.10. References

## **1.0 OBJECTIVES**

On reading this Unit you would be in a position to understand

- ❖ The concept and purpose of an Information Retrieval System (IRS)
- ❖ The functions of an IRS
- ❖ service Oriented Information Retrieval System
- ❖ The Components of an IRS
- ❖ Data Retrieval System / Data bank

## **1.1 INTRODUCTION**

Information Storage and Retrieval is often considered as the essence of Information Science or Information Studies. The term Information Retrieval was coined in 1952. The concept of Information Retrieval System (IRS) denotes a system as one that stores and retrieves information. It is composed of a set of interacting components, each of which is designed to serve a specific function for a specific purpose and all these components are interrelated to achieve certain goal. The goals are: to retrieve information in a narrower sense; to increase the level of knowledge of the users in a broader sense; to inform the users about where information is available.

According to Lancaster *“an information retrieval system does not inform (i.e. change of knowledge of ) the user on the subject of his enquiry. It merely informs him of the existence (or non-existence) and whereabouts of documents relating to his request”*.

## **1.2 THE PURPOSE OF INFORMATION RETRIEVAL SYSTEMS (IRS)**

The principle function of any library is to make the information it contains available to the library users at their request. In order to fulfill this function the information which is stored in the library must be recovered, or retrieved, from the store. The process of recovering or retrieving information is called, quite simply INFORMATION RETRIEVAL.

A library acquires documents because they contain information of the kind that is likely to be of interest to its users. The principal function of a library is information retrieval, that is the process of satisfying the requests of library users by providing them with relevant information contained within the library. The term information retrieval usually implies document retrieval i. e. the satisfaction of a request for information by retrieving a document or documents, which will contain information relevant to that request. As such it is usually distinguished from ‘ Data Retrieval’ – the satisfaction of a request for information by providing the information as a direct answer to the question. Thus we shall use the following definition of Information Retrieval: The recovery of documents from a given collection, which are relevant to a request.

By what means does a library fulfill its function of information retrieval? A library fulfils its function of information retrieval by maintaining some system for the recovery of documents from its collection. No matter how large the collection, the library is of little value if it is unable to retrieve the right documents as and when they are required. To do this it must maintain an information retrieval system. When documents to a request have been located, MATCH has been achieved between the information requested and the information retrieved. In other words, the information supplied in the document, or documents, matches, to an acceptable degree the information demanded by the user. To achieve a successful match is the central objective of information retrieval.

The possession of relevant documents, does not itself, imply a match in terms of information retrieval. To achieve a match we must be able to locate these documents within the collection. In order to locate documents relevant to a request, the collection, that is the information store, must be examined or searched. To illustrate the process of searching let us take a simple example of practical information retrieval. Suppose you wish to borrow a book from your public library about ‘programmed instruction’. Now the collection of documents in the public library covers the whole range of knowledge. In the attempt to satisfy your particular request

of information, you would not expect to have to examine every document in that collection. You would not expect to search the entire information store. You would ignore document, about history, engineering, biochemistry, physics, psychology etc., You would confine your search to those documents about 'programmed instruction, within this group of documents, probably shelved with other documents about education. You most likely find one or more relevant to your needs. At this point you would have achieved a match between information demanded and information supplied. It is obviously impracticable to search the entire information store in the satisfaction of a particular request for information. The basic principle of information retrieval is to search only a limited part of the store in response to each request, that part which is potentially relevant to the request.

The whole process of information retrieval is initiated by a request for information; now these requests are couched in a variety of ways, which express differing approaches to information needs. For example, these request for information on named subjects what has the library got on "Programmed instruction". Have you any book about classification? There are requests for a document or document by a named author or from a named publisher. These kinds of requests and many others, all exhibit valid approaches to the expression of information needs by library users. You have seen that the basic principle in Information retrieval is to search only a limited part of the store in response to each request, that part of the store, which is potentially relevant to that request. In other words, we search that class of documents which is potentially relevant to the request. It is obviously useful to have the documents themselves arranged into classes. The searcher can approach the library shelves and examine the relevant classes. If you wish to retrieve a documents arranged in classes defined by their subject content. We presumed this principle of organization in the case of searching in a public library for a document about a programmed instruction. If you want to retrieve a document by a named author, it would be useful to have the documents arranged in a classified order defined by their authorship. In fact, it would be useful to have the documents arranged into classes defined by all the characteristics by which they are sought.

Take for example, the document “ Elements of Library Classification by S. R. Ranganathan”. Let us say that this document can usefully be regarded as belonging to two classes-one defined by its authorship and the other by its subject content. It will thus receive two entries in the catalogue. One having an ‘author heading’ and the other one by a subject heading.

On searching the catalogue for information on classification this document will be indicated as relevant. This document will also be indicated as relevant on searching for documents written by S R Ranganathan Because the catalogue can contain more than one substitute for each documents, it is , said to allow for ‘Multiple Access’ to documents, that is, access via all the different characteristics by which a document is liable to be sought and by its class membership.

### **Self-Check Exercise**

#### **1. Define Information Retrieval System and state the purpose of an IRS**

**Note:**

**i). Write your answer in the space given below.**

**ii). Check your answer with the answers given at the end of this Unit.**

.....  
.....  
.....

### **1.3 FUNCTIONS OF AN IRS**

An Information Retrieval System deals with various sources of information on one-hand and user’s requirements on the other. It must:

- Analyse the contents of the sources of information as well as the user’s queries; and then
- Match these to retrieve those items that are relevant.

The major functions of an IRS can be listed according to Lancaster and Kent are:

1. To identify the information resources relevant to the areas of interest of the target user's community;
2. To analyse the contents of the sources (Documents)
3. To represent the contents of the analysed sources in a way that will be suitable for matching user's queries
4. To analyse user's queries and to represent them in a form that will be suitable for matching with the database
5. To match the search statement with the stored database
6. To retrieve the information that is relevant; and
7. To make necessary adjustments in the system based on feedback from the users.

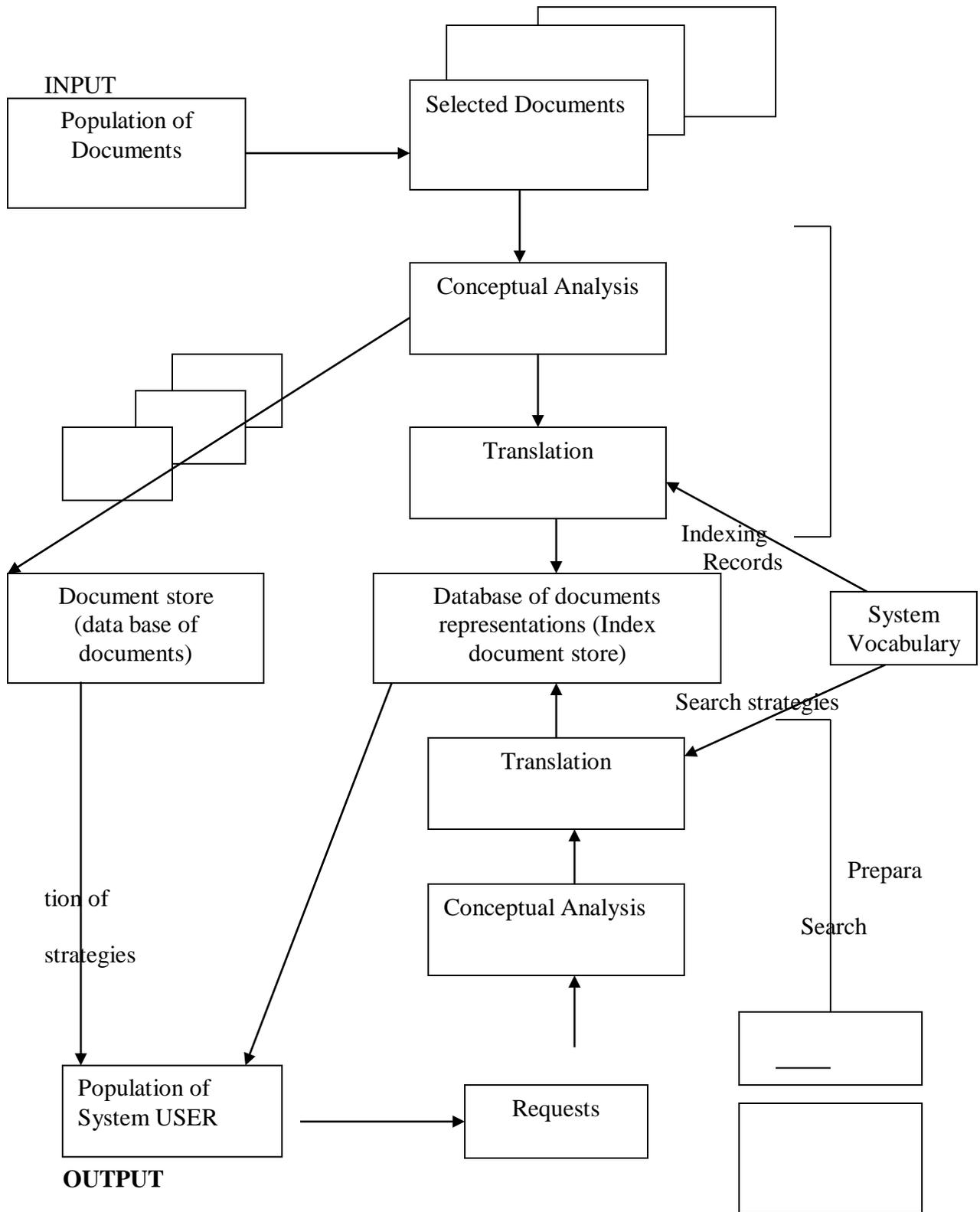
### **1.3.1 The Information Retrieval System**

The major activities of information systems are depicted in simplified form

#### **The Major function performed in many types of Information Centres**

This figure taken from F. W. Lancaster: Information retrieval system ed. 2 New York,

Wiley & Sons 1979



The system input consists of documents; that is, the information center acquires documents based on the selection criteria and policies of the library. This implies a detailed and accurate knowledge of the information needs of the user community to be served. If the documents are acquired by the systematically 'Organised and Controlled' by means of classification, cataloguing, subject indexing, abstracting and related procedures. So that they can be identified and retrieved in response to various types of user demand.

## **1.4 SERVICE ORIENTED INFORMATION RETRIEVAL SYSTEMS**

### **1.41 Question – Answering System:**

Many libraries and information centers are providing this type of question answering service. Sometimes it is referred to as a 'quick reference' service. A question answering service attempts to produce the direct answer to the particular question. For example, what is the height of Himalayas? What is the melting point of aluminium? What is the address of the British High Commission in India? This service requires reference to the particular documents that might provide the answer to the question. The first stage involves the use of some information retrieval system available in the library such as catalogue, a printed index or even an index of the book, in order to identify documents likely to provide the answer to a question posed by a user. The second stage, quite simple involves the extraction of the answer from the document and the transmission of the answer to the user.

### **1.4.2 Data Retrieval System / Data bank**

Some computer based question answering systems have been developed. Such systems accept a question in natural language and produce the answer in printed out or displayed on a screen (video display) directly, Examples are census data, thermo physical properties data or data on interaction potentials.

### 1.4.3 Passage retrieval system

A passage retrieval system is one that stores a body of text in some subject area and can retrieve a passage of text, for example, a paragraph, when it matches a search strategy representing some information need. This system is acting as an intermediate to the system that retrieves document or their representation of the user attempts to answer question directly.

#### Self-Check Exercise

#### 2. State the functions of an IRS

Note:

- i). Write your answer in the space given below.
- ii). Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

.....

.....

### 1.5 THE COMPONENTS OF INFORMATION RETRIEVAL

An information retrieval system is an organized body of knowledge having different 'components' or we might also call it sub-system-working together for a common goal or objective of the super system. No system exists in isolation, it requires an environment. The environment of the system is social that is to impart educate, entertain, instruct, the user as a whole.

A library or information center is also a system, which has to achieve its goal and also fulfill the objective of the parent body. For example a university library's super system is the university itself. The objective of the university is to promote higher

education and support research activities. The information system could either be international level, national, local, discipline oriented or mission oriented.

The library has a number of sub-systems like the reference section, circulation section, technical section, periodical section and acquisition section, etc., Each section has its own objectives. For example, the objective of the reference section is to provide the right book to the right reader at the right time.

The input components of the reference section would be the staff, reference tools and the readers' requests or queries. The purpose of reference work is to bring together the reading materials and the readers in an effective manner. So that, the right reader gets the right document at the right time. Suppose a reader requests some information the reference librarian with the help of the reference tools, process the request, that is, culls our information and provides it to the readers.

The degree of relevance of the information is ascertained from the reader and if required changes are made. This leads to the feed back mechanism, which is on later on evaluated with the help of several measures like recall and precision (Relevance)

Hence, a system is a combination of diverse but interacting elements integrated to achieve an over all objective. The elements may be electrical devices, assembly lines, or human beings concerned with processing materials, information or energy. The objective may be to guide a space vehicle, to control a chemical process or to provide a valuable service.

Each system consists of a group of logically inter-related operations. Similarly, the information retrieval system comprise, of six major sub-systems:

They are as follows

1. The document selection sub-system
2. The indexing sub-system

3. The vocabulary sub-system
4. The searching sub-system
5. The sub-System of interaction between the user and the system (user system interface) and
6. The matching sub-system

### **1.5.1 The document selection sub-system**

The library / information center acquires all materials currently published throughout the world which are of value to scholarship. This document selection sub-system involves the location, selection ordering and receipt of source materials for a collection. The process consists of a number of tasks, such as:

1. Determination of current and probable future requirements of potential users of an information retrieval system
2. Formulation of a policy of acceptance of source materials as may be defined by subject coverage, publication type, or other criteria and
3. Comparison of available or incoming source materials with policy to determine which shall be included in the information retrieval system.

### **1.5.2 The Indexing Sub-System and the Vocabulary Sub-System**

In the literature of librarianship the term indexing is used with several shades of meaning. But we are particularly interested in that aspect of indexing which supports the retrieval of documents in response to requests for information about a named subject. A system for naming subject in the way we have described is called an indexing language and like any other language, it consists of two parts; Vocabulary and syntax. If we use terms as they appear in documents without modification, we are using natural language. However, this can lead to many problems such as those arising from the use of different words by different authors to denote the same idea.

Another problem is that we can often express the same idea in more than one way using the same or similar words but altering the word order. For example, the title of document “child psychology, may be changed to ‘psychology of Children’, a ‘adult, education’ or ‘education for adults’. For these reasons, nearly all systems introduce a measure of control over the terms used, that is to say we use a controlled vocabulary, vocabulary control involves the establishment of relationships among analytic, often on an arbitrary basis. But most usually based on the prediction of those relationships that may facilitate identification of all source materials that have been indexed. This flexible syntax of natural language is formalized to permit only certain constructions. For example, instead of children’s libraries and libraries for children we use libraries, children’s libraries and libraries for children we use libraries, children’s. A controlled vocabulary is part of an artificial indexing language. The extreme example of an artificial language is the notation of a classification scheme; instead of the natural language terms heat treatment of aluminium or aluminium, heat treatment, we use 669.71’04.

Almost all classification schemes are controlled by a thesaurus. Thesaurus is a book of words, that shows explicitly the relationships among the words it contains. These relationships may be those of

- : Synonymous ;
- : Specific to generic (often called Broader term)
- : Generic to specific (often called Narrower term)
- : General non-specific relationship (Often called related term)

The most famous example is “Roget’s Thesaurus of English words and Phrases.

### **1.5.3 The searching Sub-System**

So far we have considered the passive approach; the library provides materials, which the readers select in a fairly indeterminate way. However, much of the use

made of libraries is active: readers come to the library seeking information on particular subjects, and expect our system to be able to provide the answers. The basic steps that must be taken, in conducting any searching operations are as follows:

1. A question or problem must exist and be recognized and must be verbalized or recorded for communication to the search system;
2. The question must be analyzed in order to select analytics (or Clues) that will be useful in formulating a strategy of search;
3. The analytics selected must be transformed into a language and into a strategy configuration that conforms to those of the system used for analysis and storage of the records of the file.
4. The analytics and search strategy selected must be formalized in terms of a language and program that will conform to those of the device used for searching.
5. The searching machinery must be set in motion and
6. The response must be obtained.

#### **1.5. 4 The Sub –System of interaction between the user and the system (User Systems interface)**

In this sub-system of interaction between the user and the system, the receiver becomes a source, encoding a message in the form of an enquiry. We now have to discover any information in our store which appears to match the enquiry and we can pass them on to the enquirer, who can decide whether they match his needs. In the light of our response, the enquirer may modify his message to match his requirements.

#### **1.5.5 The Matching Sub-System**

The sub-system that actually matches documents representations against request representation that is when documents relevant to a request have been located, a match

has been achieved. In other words, the information supplied in the document, or documents, matches to an acceptable degree to the information demanded by the user. In a conventional computers based system the computer based system the computer contributors directly only to the matching operation. It acts as a giant matching device. The matching sub-system has no direct influence on the effectiveness of the complete system, that is, on whether or not it can retrieve items that satisfy the information needs of users, although, clearly, the efficiency of the matching sub-system exerts a rather great influence on system economics and over all system efficiency, measured for example, in response time.

## 1.6 Check your progress

1. Who coined the phrase 'Information Retrieval'?

(A) **Calvin Mooers** (B) S.R.Ranganathan (C) J.D.Brown (D) H.P.Luhn

2. Information retrieval system is.....

Ans:- According to Lancaster "an information retrieval system does not inform (i.e. change of knowledge of ) the user on the subject of his enquiry. It merely informs him of the existence (or non-existence) and whereabouts of documents relating to his request".

3. "Anomalous State of Knowledge (ASK) Model", one of the user centered models of information retrieval was proposed by

(A) P.Ingwensen (B) T.Saracevic (C) **J.N.Belkin** (D) D.Ellis

4. The 'Cranfield Test', associated with evaluation of information retrieval system was carried under the direction of \_\_\_\_\_

(A) **C.W.Cleverdon** (B) G.Salton (C) D.C.Blair & M.E.Maron (D)

E.Voorhees

5. Which law states that "An information retrieval system will tend not to be used whenever it is more painful and troublesome for a customer to have information than for him not to have it"?

- (A) Bradford's Law
- (B) Mooer's Law**
- (C) Ziff's Law
- (D) Lotka's Law

### **1.7. SUMMARY**

This introduces the concept of Information Retrieval system, which denotes a system as one that stores and retrieves information. It is composed of a set of interacting components, each of which is designed to serve a specific function for a specific purpose and all these components are interrelated to achieve certain goal. An IRS is designed to retrieve the documents or information required by the and user community. It should make the right information available to the right user. Thus an IRS aims to collect and organise information in one or more subject areas in order to provide to the required clients. An IRS serves and acts a mediator or a bridge between the information generators or creators and the information users.

### **1.8. Glossary**

**Information Retrieval:** The location and presentation to a user of information relevant to an information need expressed as a query

**Information Retrieval System:** Any system usually involving computers, that performs information retrieval.

**System:** A system is a whole or unity comprised of interrelated and interdependent parts (subsystems) . The whole performs some functions to which each of the subsystems contributes by carrying some specific operations directly or indirectly related to the function of the whole.

### **Questions for self study**

- 1. Define Information Retrieval System and state the purpose of an IRS**

2. State the functions of an IRS
3. State the components of an IRS

### **1.9.REFERENES AND FURTHER READING**

Chowdhury, G. G .1994 Information Retrieval System. Calcutta: IASLIC.

Chowdhury, G. G .1999 Introduction to modern Information Retrieval System. London: Library Association Publishing.

Lancaster, F. W. 1979. Information Retrieval Systems: Characteristics, testing and evaluation. 2<sup>nd</sup> ed. New York: John Wiley and Sons.

## **UNIT - 2: SEARCH STATEMENT**

### **Structure**

2. 0 Objectives

2.1 Introduction

2.2 Basic Searching

2.3 Building A Search Statement

2..4 Major Search Formulations

2.5 Search Strategy And Its Pre-Requisites

2.6 Evaluation And Feedback

2.7 Factors Affecting The Success Or Failure Of A Particular Search

2.8 Check your progress

2.9 Summary

2.10Glossary

2.11Questions for self study

2.12References

## 2.0 Objectives

After reading this Unit you will be in a position to understand:

- ❖ The concept of Basic Searching and how to Building a Search Statement;
- ❖ Major Search Formulations
- ❖ Search Strategy and its Pre-Requisites;
- ❖ Evaluation and Feedback in the search and
- ❖ Factors affecting the Success Or Failure of a Particular Search

## 2.1 INTRODUCTION

The search proper begins after the user has logged on to the search service. On some hosts log-on followed by news from the host about such matters as amendments to the list of databases available, and changes the command language or network protocols. Once any news items have been displayed, the user will be prompted by the host to enter the first step in the search strategy. Search in an IRS is essential a heuristic process. The overall search strategy that might be employed in the course of a search are influenced not merely by the end user's requirements but also by the characteristics of the IRS, the devices it offers, etc.

The concept of online searching has occupied a large and significant area in the study and research of modern information retrieval.. The cost of searching a database can be reduced only if, an appropriate search strategy is followed. There are many issued are to be considered while formulating a search statement such as:

- The concepts or facets to be searched and their order;
- The term(s) that appropriately represents the search concept
- The features of the retrieval system concerned;
- The measurers to be taken in revising a search statement

To formulate a query, a user must select collections, metadata descriptions, or information sets against which the query is to be matched, and must specify words, phrases, descriptors, or other kinds of information that can be compared to or matched against the information in the collections. As a result, the system creates a set of documents, metadata, or other information type that match the query specification in some sense and displays the results to the user in some form.

## **2.2 BASIC SEARCHING:** Command Driven systems;

Any online information retrieval system must provide some method for the searcher to issue instructions to the computer and in return to receive messages back from it. The searcher may want to issue an instruction to look into a particular database, to find out which records contain one or more terms or match a particular characteristic (language, type of publication etc.,) or to display records on a screen or on a print-out.

In some cases this dialogue between the searcher and the computer is conducted through a series of menus, which present the searcher with choices from which a selection must be made. Alternatively, interaction between the searcher and the computer may be command driven. This approach is still more commonly encountered in online search services than the menu approach.

### **2.2.1 Pre-search preparation**

In the last two decades bibliographic search systems have become extremely sophisticated. The number of databases available, their size and the differences in indexing systems have become considerably simpler and more comprehensive today when compared to their pioneer services in the 1970s. There were very few databases and their size was relatively small. The users were having very limited options. Users need to enter strings of keywords, or sets of single keywords, and combined them with Boolean operators. The ability to search one database and execute it in another database was unknown. Users had no option but to re-enter the search in each database to be searched.

Today, the scenario of database has completely changed with regard to the number of database available, their size and the search systems they offer. The search system software capabilities have grown to facilitate efficient database selection and searching with a minimum effort, say by a click of the mouse. The same search can be carried in a number of databases without re-entry of the search options. Users can even store their searches offline for further refinement for later execution. These developments cut down on the time a user must spend online to retrieve the desired information. However, there is one thing that the users cannot escape the database searching systems – building the search strategy.

**Self-Check Exercise**

**1. What are the issues to be considered while formulating a search statement?**

**Note:**

**i). Write your answer in the space given below.**

**ii). Check your answer with the answers given at the end of this Unit.**

.....  
.....  
.....

**2.3 BUILDING A SEARCH STATEMENT**

Search strategy encompasses all activities involved with search a database right from the reference interview with the user to the verification of final output. The knowledge of databases and their search strategy methods are essential for a librarian as well as end users.

**2.3.1 Steps in building a search statement**

An IRS and a searcher who may be the need user or the information intermediary, the process of information retrieval or the search process can be broken down into discrete steps, which are generally carried out in that order during the retrieval process. These steps are:

1. Understanding the information needs of the end user;
2. Formulating search objectives;
3. Selecting one or more appropriate databases relevant to the information need (as many IR operations take place in computer-based systems it may be necessary to select a search system / database vendor);
4. Identify the major concepts and their interrelationships and ways of expressing these concepts (words / phrases, etc);
5. Identify the fields of the database records to be searched (Title field, descriptor field, abstract field, etc.)

6. Translate decisions made above into formal search statements using the command language of the search system;
7. Connect to the search system, call the appropriate database(s) and execute the search formulated;
8. Examine the initial search output and take decision regarding continuing the search or modifying the search or reformulating the search. The search process is generally continued until reasonably satisfactory output is obtained.

## **2..4 MAJOR SEARCH FORMULATIONS**

Every search has certain goals in terms of the requirements of the end user. Search Strategy may be thought of as the overall plan for realising these search objectives. The term search tactic applied primarily in the context of online searching refers to the heuristics employed to advance a particular strategy. The major search strategies that have been reported in the literature on online searching include:

- Brief search;
- Building blocks search strategy;
- Successive Facet Strategy; etc.

There are also certain strategies that can be employed in a citation index. The employment of a strategy should be decided in terms of its ability to achieve the objectives for which the search is being carried out. It is not our objective here to give a detailed explanation of the various search objectives, as these are best understood in a real life situation. However, a general idea of these strategies, which can be employed in a search, is in order.

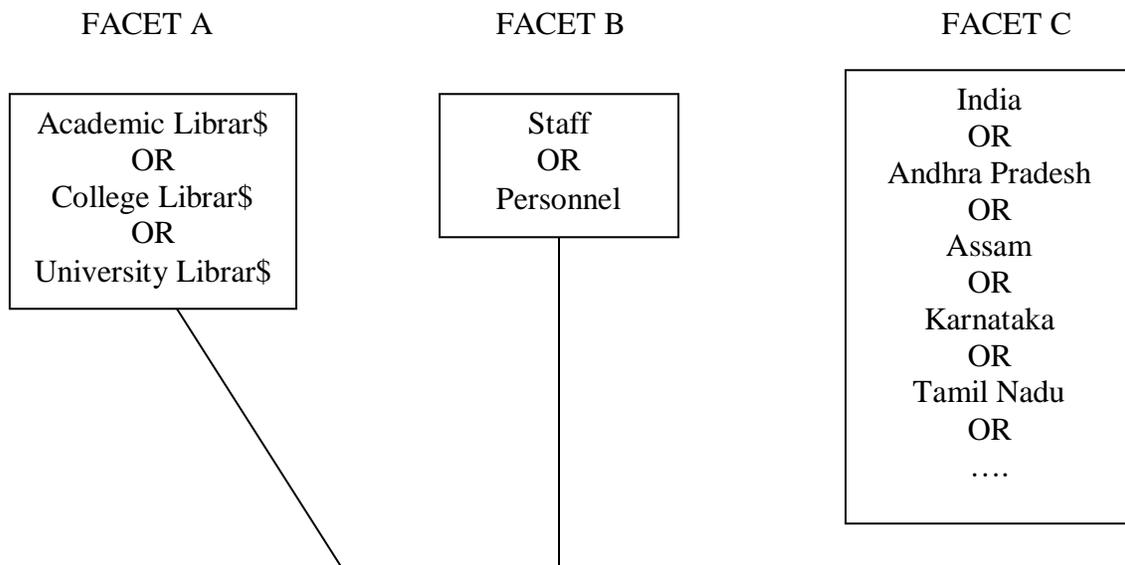
### **2.4.1 Brief search**

A Brief search as the name suggests is a single search formulation, usually a Boolean combination of search terms. The purpose invariably is to retrieve a few relevant documents. Often such a search is exploratory in nature and the result may be employed to formulate a more detailed and effective search. For example, the retrieved documents may be used for the purposes of identifying additional search descriptors.

## 2.4.2 Building Blocks Search

The Building Blocks Search which is probably the most widely employed search strategy in online searching involves, first, the identification of every important concept of the search (query) which are generally grouped together under two / three facets. For each facet the possible search terms are identified and are connected using the Boolean OR to constitute the building blocks. These blocks are combined again using appropriate Boolean operators to formulate the search. Let us consider an example: Suppose the end user is interested in identifying documents that deal with 'Staff in academic libraries in India', the building blocks strategy will proceed as below. There are three concepts / facets in this query, viz. 'India', 'Academic Libraries' and 'Staff'. We need to select terms to represent each of these facets, which may be synonyms of the term or narrower terms or, in some cases, even related terms.

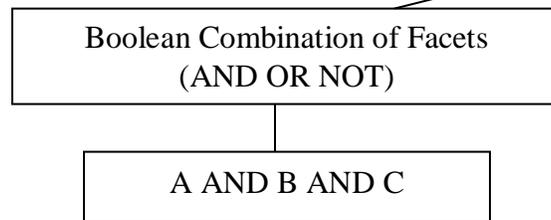
In the course of the search heuristics are employed to increase Recall (By adding via OR additional concepts) or precision (by deleting a few broader concepts or ambiguous terms) as may be required by the end user.



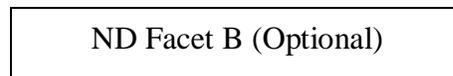
## Fig. Building Blocks Search

### 2.4.3 Successive Facet Strategy

Successive Facet Strategy employs building blocks but, to begin with, search is

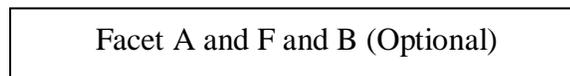


not conducted in all the facets. Facets are constructed



AND

**FACET C (Optional)**



AND

**FACET C (Optional)**

#### Self-Check Exercise

**2. How do you build a search statement?**

**i). Write your answer in the space given below.**

**ii). Check your answer with the answers given at the end of this Unit.**

.....  
.....  
.....

## **2.5 SEARCH STRATEGY AND ITS PRE-REQUISITES**

Search strategy encompasses several steps and levels of work in information retrieval. Meadow mentions that search strategy includes at least three decision points, which a searcher has to reach. There are many questions, which need consideration in formulating an appropriate search statement, like-

1. The concepts are facets to be searched and their order;
2. term(s) which appropriately represents the search concept;
3. the feature(s) of the retrieval system to be approached;
4. the measures to be taken in revising a search statement; and so on.

Developing a good search strategy requires knowledge about the nature and organisation of target database(s) and also the exact need of the user. Understanding of the user's exact requirement has much impact on the actual search and retrieval process. In some cases, the user may only want some relevant items on a given topic, in which case the task of searching will obviously be limited. Conversely, the user may want to get all relevant items (obviously with as less number of non-relevant items as possible), in which case the search must be exhaustive. Meadow identifies three kinds of search, viz.,

1. High recall search: When the user needs to find out all the relevant items on the stated topic;
2. High precision search: When the user needs only relevant items, i.e. as less number of non-relevant items as opposed to all relevant items.

## **2.6 EVALUATION AND FEEDBACK**

When a search is completed and an output is obtained the searcher should necessarily obtain the user's assessment of the search results. From the point of view of the end user it is likely that the evaluation will be based on how well the results satisfy his/her information needs. It is also important to assess the retrieval effectiveness and to analyse the factors contributing to search failure.

## **2.7 FACTORS AFFECTING THE SUCCESS OR FAILURE OF A PARTICULAR**

## SEARCH

The various factors affecting the success of a search have been mentioned below. The factors most directly related to the search subsystem itself may be summarised as follows:

1. The searcher's interpretation of the needs of the user. The prime factor is the quality of the interaction between the requester and the system
2. Given a request statement that closely matches the requester's needs, one factor that influences the search results is the complexity of the request. The "simpler" the request, that is, the fewer facets involved, the better the result is likely to be.
3. Quite certainly, the performance for any request depends on the ability of the index language to express precisely the concepts involved in the request. The vocabulary of the system must be capable of expressing the subject matter of the request at a reasonable level of specificity.
4. In any particular information system, there may be certain subject areas in which the performance, on the average, is likely to be worse than that in other subject fields.
5. Indexing policies and practices affect the performance level that can be achieved in a particular search.
6. The capabilities of the searching software in use in the system also exert some influence on the performance of the search, for this governs just what the searcher is or is not able to do – for example, whether or not he can truncate terms.
7. A search can be ruined or substantially reduced in value by an inadequate or inaccurate strategy.

### 2.8. Check your progress

1. To carry out a search on 'Poverty in Gujrat and Rajasthan' a search statement would need to be framed as

**(i) Poverty AND (Gujrat AND Rajasthan)** (ii) Poverty AND (Gujrat OR Rajasthan) (iii) Poverty OR (Gujrat AND Rajasthan) (iv) Poverty OR (Gujrat OR Rajasthan).

2. Queries of the users are translated into the indexing system and matching is done with the vocabulary of the system

- (a) Query formulation
- (b) Query assimilation
- (c) Query matching
- (d) Search strategy**

3. According to George Boole, three types of Boolean searching are

- (a) AND, BUT, NOT
- (b) AND, NOR, BUT
- (c) AND, OR, NOT**
- (d) AND, OR, BUT

4. What is Thesaurus?

- (a) A collection of selected terminology**
- (b) Synonym terms
- (c) List of words
- (d) All of the above

5. 'Recall' and 'Precision' are the terms used in

- (i) reference service
- (ii) information retrieval system**
- (iii) library management
- (iv) book selection.

## **2.9 SUMMARY**

The concept of online searching has occupied a large and significant area in the study and research of modern information retrieval. The cost of searching a database can be reduced only if; an appropriate search strategy is followed. Search strategy encompasses all activities involved with search a database right from the reference

interview with the user to the verification of final output. The knowledge of databases and their search strategy methods are essential for a librarian as well as end users. The major search strategies that have been reported in the literature on online searching include: Brief search; Building blocks search strategy; Successive Facet Strategy; etc.

There are many questions, which need consideration in formulating an appropriate search statement, like-the concepts are facets to be searched and their order; term(s) which appropriately represents the search concept; the feature(s) of the retrieval system to be approached; the measures to be taken in revising a search statement; and so on. There are various factors affecting the success of a search, which are most directly related to the search subsystem itself.

## **2.10 Glossary**

**Search Statement:** Instruction to the computer to find records matching the term or combination of terms entered by the searcher.

**Search Strategy:** The plan for how a request will be searched on the computer. It will include a series of search statements combined by Boolean operators, which will normally be planned in advance.

## **2.11. Questions for self study**

1. What are the issues to be considered while formulating a search statement?
2. How do you build a search statement?

## **2.12 REFERENCES**

Walker, Geraldene and Janes, Joseph. 2005. Online Retrieval: A dialogue of Theory and Practice. Colorado: Division of Greenwood Publishing group.

Chowdhury, G. G .1999 Introduction to modern Information Retrieval System. London: Library Association Publishing.

Hartley, R. J. et al. 1993. Online searching: Principles and Practice. London: Bowker-Sour.

Lancaster, F. W. 1979. Information Retrieval Systems: Characteristics, testing and evaluation. 2<sup>nd</sup> ed. New York: John Wiley and Sons.

Walker, G and Janes, J 1999. Online Retrieval: A dialogue of theory and practice. 2<sup>nd</sup> ed. Colorado: Libraries Unlimited.

## **UNIT – 3: DIFFERENT TYPES OF QUERY FORMULATION**

### **Structure**

3.0 Objectives

3.1 Introduction

3.2 Query Structure

3.3 Structured Query Formulation

3.4 Query Language

3.5 Faceted Queries

3.6 Relation Statistics

3.7 Check your progress

3.8. Summary

3.9. Glossary

3.10. Questions for self study

3.11. References

### **3.0 OBJECTIVES**

After reading this Unit you will be able to:

- Know the concept of Query language,
- Query structure and Structured Query Formulation(SUF);
- Understand the methods of creating queries in a structured manner; and
- Acquaint yourself with the different methods and techniques of query optimization.
- Relation Statistics

### **3.1 INTRODUCTION**

The utilities of query languages are for database manipulation. There are different query languages, which may be used, depending upon the nature of the database management system (DBMS). A few of them are ISBL, SEQUEL, SQUARE, Query-By-Example, SQL, AQL etc. As you know, the relational database management system (RDBMS) is a view of DBMS and is widely accepted. The details of RDBMS have been detailed in Block 3 of Information systems. Structured query language (SQL) is the well-accepted query language along with RDBMS. For searching in Internet, new query languages for HTML, XML documents are evolving. XML-QL is one among them to mention.

### **3.2 QUERY STRUCTURE**

The language of the query may be quite constrained, indeed a query may not satisfy the normal rules of syntax. The document and the query undergo parallel processes within the retrieval system. On the document side, someone generates or gather some data and formulate it into a document. From the end user side the document is transformed into internal representation. Similarly on the query side someone begins with an information need, using it to generate a query. A good system design will make these transformations as simple and automatic as possible, largely transparent to the user.

### **3.3 STRUCTURED QUERY FORMULATION**

(SQL) is not a complete programming language. It is used to interrogate and process data in a relational database environment. SQL commands can work interactively with a database or be embedded in programming languages such as C, COBOL or Java. IBM developed SQL for using in mainframes.

SQL is a non-procedural, fourth generation language that allows users to access data in relational database management systems, such as Oracle, Sybase, Informix, Microsoft SQL Server, Access, and others, by allowing users to describe the data that he or she wishes to see. SQL also allows users to define the data in a database, and manipulate that data. This unit will describe how to use SQL, with examples. The SQL is in both ANSI and ISO standards. The SQL described in this unit is standard SQL. Some SQL features of specific database management systems are mentioned in the Nonstandard SQL.

Structured Query Formulation (SQF) is the method of creating queries in a structured manner in order to get the results accurately and efficiently. Every database query statement works on a database, and requires reading and writing of database records or data blocks. It is necessary to understand how these queries work, and how many read and write is involved with each query, etc. It is possible to get the same result by reducing the read/wrote operation, by changing the query a little differently sometimes or by formulating the query in a more structured manner. The procedure of analyzing a query and converting it into another query statement of the same meaning is known as query optimization. Structured query formulation is possible by understanding the different methods and techniques of query optimization. A structured query is a result of a query optimization procedure on a non-structured query. This unit will help you understand query optimization and hence to write more efficient and structured queries.

Database queries have been wrist ten by using and query language. The flexibility, user friendliness and easiness of this language that one should only know what is the result expected after executing the query. And one need not know how the data is arranged, how the query is executed in terms of reads and writes of database blocks, etc. As a result the way the query is written, though it gives the result, it need not be efficient

and fast in executing the query. It is very important to know how the query works and how a more optimized and efficient query of the same meaning can be written. Query optimization means to convert a query  $Q$  to its semantically equivalent query  $Q'$ , which evaluates faster than  $Q$ . The newer Database Management Systems (DBMSs) provide/support declarative query languages, using such a query language. Users express their queries in terms of “what” they want rather “how” to obtain it; thus, the extra burden of finding a good access path and an efficient evaluation strategy is left to the DBMS.

Query optimization is concerned with the techniques used by a DBMS to reduce the execution time of a query. All of the major database vendors now include optimizing SQL compilers, which analyses the SQL query sent to it, rewrites the query if necessary, and finally produces an optimal access plan to retrieve the data from the database. This module of SQL compiler is called Query Optimizer. This optimization is done based on the different optimization rules formulated on the basis of the cost of each operation on the query. However, regardless of the improvements made in query compiler technology, the implementer will have a much better knowledge of how the database is constructed and how the applications interact with this database. In addition, optimizers can and often do make the wrong decision.

While SQL optimization is a very important part of the process of finely tuning a database application, it should not be considered in isolation since there are a number of other areas, which also require careful attention from the implementer. Some examples of areas which should also be scrutinised are: overall database design and structure, indexing, table space structures and types, database configuration parameters, hardware issues like memory, processor speed, and disk types and so on.

This unit deals with the cost implication of each query operation and how to optimize the total cost of a query. It also gives some optimization strategies to be followed. To understand these optimization methods, it is required to know the relation statistics.

### **Self-Check Exercise**

- 1. Explain the concept of Query structure.**

**2. What do you mean by Structured Query Formulation?**

**Note:**

**i). Write your answer in the space given below.**

**ii). Check your answer with the answers given at the end of this Unit.**

.....  
.....  
.....

**3.4 QUERY LANGUAGE**

The users, instead of querying the target collection and going through multiple iterations of the process (query formulation,, ,,query evaluation, , relevance assessment), will first interact with the source collection to formulate a query, which is then issued to the target collection to retrieve the relevant documents available there. This becomes the query language of the day. To formulate a query, a user must select collections, metadata descriptions, or information sets against which the query is to be matched, and must specify words, phrases, descriptors, or other kinds of information that can be compared to or matched against the information in the collections. As a result, the system creates a set of documents, metadata, or other information type that match the query specification in some sense and displays the results to the user in some form.

Shneiderman identifies five primary human-computer interaction styles. These are: *command language, form fillin, menu selection, direct manipulation, and natural language*. Each technique has been used in query specification interfaces and each has advantages and disadvantages. These are described below in the context of Boolean query specification.

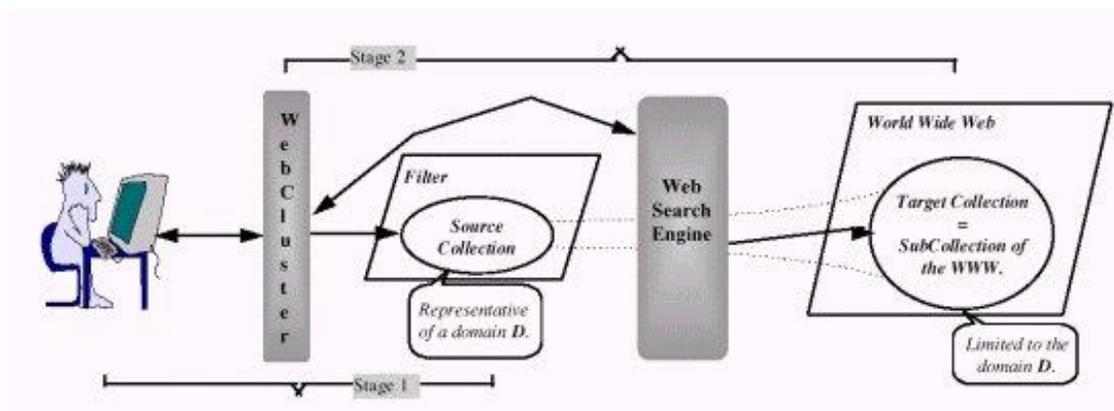
Therefore, searching the becomes a two stage process:

1. The Query formulation stage helps the users to formulate a **precision oriented** query. They search (browsing/querying) the structured source document collection to gain knowledge about the inherent semantic structure of the domain covered by it and retrieve some relevant source documents or cluster of documents. This stage will allow them to learn the important concepts of the

domain and the ones corresponding to their information need, and how they are represented in this document collection. At the end of this stage users should have identified their information need either as a concept (or set of concepts) of the specific domain, or as a cluster of relevant documents.

2. The *Mediated Access* stage uses the information need expression identified in the previous stage to issue a query to the target collection, and allow the users to browse through the result hence produced.

It is believed that in the need for some sort of intelligent assistance to filter the information available on the WWW and reduce it to specialized subsets, which were selected according to the *domain of interest* of a particular user or group of users. This assistance should also provide some guidance in the query formulation stage, by helping the user in building an effective query which is representative of its information need and precise enough to exclude from the result most of the non-relevant documents. The Fig.-1 shows the Mediated Access model.



**Fig. 1 Mediated Access model**

Perhaps the most common operation on any database is retrieval of information. Therefore, one needs efficient and accurate methods for retrieving data. Queries can be formulated using several techniques. These techniques fall broadly into two categories: textual and visual. Text can be used to formulate queries for visual data (images, video,

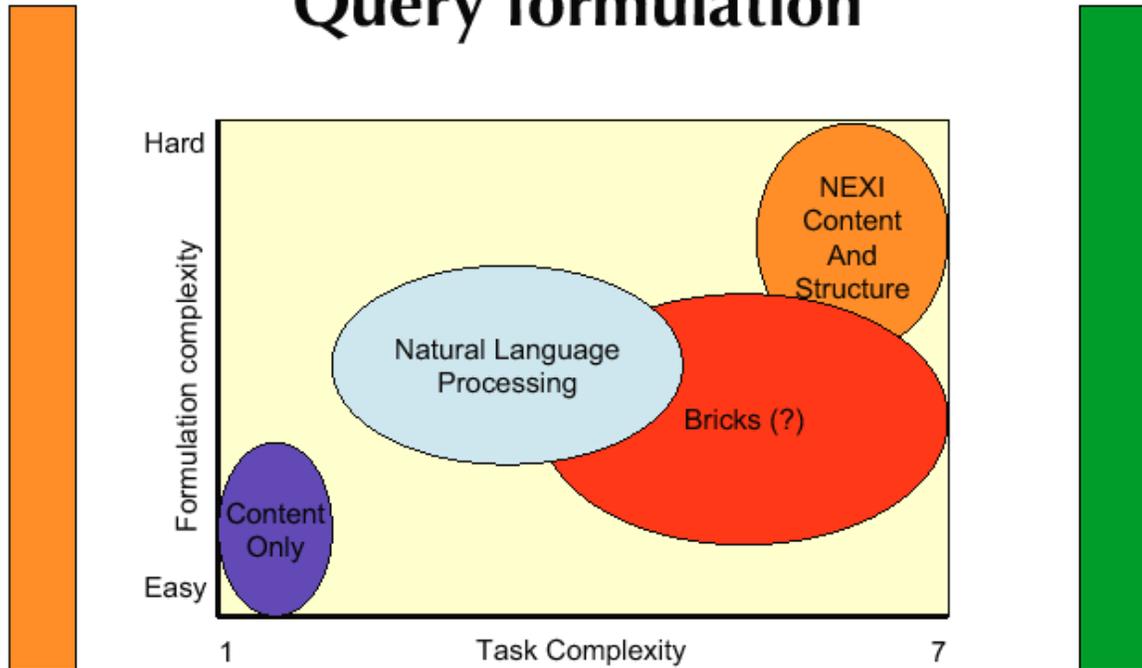
graphs), but such queries are not very efficient and cannot encompass the hierarchical, semantic, spatial, and motion information.

The interactive query process consists of three basic steps: *formulating the query*, *processing the query*, and *viewing the results returned from the query*. This requires an expressive method of conveying what is desired, the ability to match what is expressed with what is there, and ways to evaluate the out come of the search. Conventional text-based query methods that rely on keyword look-up and string pattern-matching are not adequate for all types of data, particularly auditory and visual data. Therefore, it is not reasonable to assume that all types of multimedia data can be described sufficiently with words alone, neither as meta-data when it is first entered in the database, nor as queries when it is to be retrieved.

Visual data can also be very difficult to describe adequately by keywords as each user based on his or her impression of the image and video chooses keys. Thus it is difficult, if not impossible to know under what keyword the target has been indexed. In addition, keywords must be entered manually, are time consuming, error prone, and thus cost prohibitive for large databases. A query system that allows retrieval and evaluation of multimedia data should be highly interactive to facilitate easy construction and refinement of queries. Due to the visual nature of the data, a user may be interested in results that are similar to the query, thus, the query system should be able to perform exact as well as partial or fuzzy matching. The results of a query should be displayed in a decreasing order of similarity, from the best match to the  $n^{\text{th}}$  based matched result for easy browsing and information filtering. Due to the complex nature of video queries, there should be a facility to allow a user to construct queries based on previous queries.

Query Formulation method is shown in Figure - 2

# Query formulation



Outline - **Motivation** - Objectives - Theory - Usability - Conclusions

Centre for Content and Knowledge Engineering, Utrecht University, the Netherlands

## Self-Check Exercise

### 3. Define Query language and Explain Query formulation

Note:

- i). Write your answer in the space given below.
- ii). Check your answer with the answers given at the end of this Unit.

.....  
 .....  
 .....

### 3.5 FACETED QUERIES

A typical problem with Boolean queries is that their strict interpretation tends to yield result sets that are either too large, because the user includes many terms in a disjunct, or are empty, because the user conjoins terms in an effort to reduce the result set. This problem occurs in large part because the user does not know the contents of the collection or the role of terms within the collection.

A common strategy for dealing with this problem, employed in systems with command-line-based interfaces like DIALOG's, is to create a series of short queries, view the number of documents returned for each, and combine those queries that produce a reasonable number of results. For example, in DIALOG, each query produces a resulting set of documents that is assigned an identifying name. Rather than returning a list of titles themselves, DIALOG shows the set number with a listing of the number of matched documents. Titles can be shown by specifying the set number and issuing a command to show the titles. Document sets that are not empty can be referred to by a set name and combined with AND operations to produce new sets. If this set in turn is too small, the user can back up and try a different combination of sets, and this process is repeated in pursuit of producing a reasonably sized document set.

This kind of query formulation is often called a *faceted* query, to indicate that the user's query is divided into topics or facets, each of which should be present in the retrieved documents. For example, a query on drugs for the prevention of osteoporosis might consist of three facets, indicated by the disjuncts

(osteoporosis OR `bone loss')

(drugs OR pharmaceuticals)

(prevention OR cure)

This query implies that the user would like to view documents that contain all three topics.

A technique to impose an ordering on the results of Boolean queries is what is known as *post-coordinate* or *quorum-level* ranking. In this approach, documents are ranked according to the size of the subset of the query terms they contain. So given a query consisting of `cats,' `dogs,' `fish,' and `mice,' the system would rank a document with at least one instance of `cats,' `dogs,' and `fish' higher than a document containing 30 occurrences of `cats' but no occurrences of the other terms.

Combining faceted queries with quorum ranking yields a situation intermediate between full Boolean syntax and free-form natural language queries. An interface for specifying this kind of interaction can consist of a list of entry lines. The user enters one

topic per entry line, where each topic consists of a list of semantically related terms that are combined in a disjunct. Documents that contain at least one term from each facet are ranked higher than documents containing terms only from one or a few facets. This helps ensure that documents which contain discussions of several of the user's topics are ranked higher than those that contain only one topic. By only requiring that one term from each facet be matched, the user can specify the same concept in several different ways in the hopes of increasing the likelihood of a match. If combined with graphical feedback about which subsets of terms matched the document, the user can see the results of a quorum ranking by topic rather than by word. This idea can be extended yet another step by allowing users to weight each facet. More likely to be readily usable, however, is a default weighting in which the facet listed highest is assigned the most weight, the second facet is assigned less weight, and so on, according to some distribution over weights.

### **3.6 RELATION STATISTICS**

When we try to determine if one query will be evaluated faster than another, it depends mainly on the number of disk block reads and writes. A block is the smallest amount of data that the disk hardware can read. A block, in general, is 1024 bytes. So a single block usually contains quite a few tuples of a database. Generally the disk I/O is slower than memory I/O. Relations are normally large and cannot reside entirely in memory; they must be read in and write out of memory during a query. Also, a database query will usually perform a very simple in-memory computation, so in general we can completely ignore the in-memory cost of a query.

It is necessary to know some important statistical parameters of a query for analyzing the same. For a relation,  $R$ , the following statistics are required to optimize the query.

card  $R$                     the cardinality or number of tuples in  $R$

degree  $R$                     the number of attributed in  $R$

blocks R the number of blocks that R occupies on disk

values(R, A) the number of different values for attribute A in relation R. We will assume that the A values are uniformly distributed in R, that is every different value appears in the same number of tuples

clusters(R,AS) if R is clustered, the size of a cluster, for a single value for attribute A, is the same as blocks R/values(R,A) (assuming that the values are uniformly distributed)

Consider a student relation with name and IDNo. And suppose the information is stored in blocks as below.

Block Number	1	2	3
Tuples	Rishi Gupta, AISO2001, Sunil satpal, AISO2008,  Deepika chaturvedi, AISO1004	Sunil Satpal, AISO3008, Deepika Chaturvedi, AISO2004  Rishi Gupta, AISO1001	Deepika Chaturvedi, AISO3004 Rishi Gupta, AISO3001  Sunil Satpal, AISO3008

This relation has the following statistics

card(student) = 9

degree(Student) = 2

blocks(student) = 3

values(student, name) = 3

clusters(student, name) = 1, since the relation is un-clustered and un-indexed on name.

When the relation is indexed on name, the tuples in the relation with the same name attribute are stored in the same or adjacent blocks. In other words, the same values are clustered together in the same area on disk.

Block Number	1	2	3
Tuples	Deepika chaturvedi,AISO3004 Deepika chaturvedi,AISO2004 Deepika chaturvedi,AISO1004	Deepika chaturvedi,AISO2001 Deepika chaturvedi,AISO1001 Deepika chaturvedi,AISO3001	Sunil Satpal, AISO2008 Sunil satpal, AISO3008 Suni; Satpal, AISO3008

R is indexed on attribute Name. An index is a special data structure that permits rapid search for tuples on a specific attribute (Name). Without an index, on average, we would have to read half of the relation (block by block) to find a tuple with a particular attribute value. But with an index we can go directly to the block or blocks that contains the desired tuple. It is also easier to find all the tuples with the same attribute since all of them are clustered in the same location. Indexing also involve some initial cos. But it can be ignored compared to the cost involved in searching.

There are two types of indexes-clustered index, and non-clustered index. In clustered index the data is physically sorted, while a non-clustered index is a separate index structure independent of the physical sort order of the data in the table.

### 3.7. Check your progress

1. 'Truncation' is a technique used in

- (i) cataloguing
- (ii) classification
- (iii) reference service
- (iv) online information retrieval.**

2. The logic 'And', 'Or', and 'Not' was devised by

- (i) F. W. Lancaster
- (ii) George Boole**

(iii) S. R. Ranganathan

(iv) M. A. Gopinath.

3. The query on “Use of Audio Visual Aids in teaching Science at Primary Schools” can be framed as

(i) **Audio Visual aids and Science and Primary Schools**

(ii) Audio Visual aids or Science or Primary Schools

(iii) Audio Visual aids or Science and Primary schools

(iv) Audio Visual aids and Science or Primary Schools.

4. Structured Query Formulation (SQF) is.....

Ans:- The method of creating queries in a structured manner in order to get the results accurately and efficiently.

5. A Boolean query using only AND and NOT is

(A) Fuzzy query

**(B) Conjunctive query**

(C) Routing query

(D) Probabilistic query

### **3.8 SUMMARY**

In modern information access systems the matching process usually employs a statistical ranking algorithm. However, until recently most commercial full-text systems and most bibliographic systems supported only Boolean queries. Thus the focus of many information access studies has been on the problems users have in specifying Boolean queries.

### **3.9 Glossary**

**Query:** The formal expression of an information need.

**Query Language:** The expression of the user information need in the input language provided by the information system. The most common type of input language simply allows the specification of keywords and of a few Boolean connectives.

### **3.9. Questions for self study**

1. Explain the concept of Query structure.
2. What do you mean by Structured Query Formulation?
3. Define Query language and Explain Query formulation

### **3.10 References**

Baeza-Yates, Ricardo and Riberiro-Neto, Berthier. 2004 Modern Information Retrieval. Patpargan: Pearson Education

Chowdhury, G. G .1994 Information Retrieval System. Calcutta: IASLIC.

Korfhage, Robert R. 1997 Information Storage and Retrieval. New York: John Wiley & Sons.

Chowdhury, G. G .1999 Introduction to modern Information Retrieval System. London: Library Association Publishing.

Lancaster, F. W. 1979. Information Retrieval Systems: Characteristics, testing and evaluation. 2<sup>nd</sup> ed. New York: John Wiley and Sons.

## **UNIT- 4: SEARCHING AND SEARCH PROCESS**

### **Structure**

4.0 Objectives

4.1 Introduction

4.2 Aims of search strategy

4.3 Basic Principles Of Search Strategy Formulation

4.4 Types Of Search Strategies

4.5 Preparation Of Search Strategy

4.6 How To Search And What To Search

4.7 Check your progress

4.8 Summary

4.9 Glossary

4.10 Questions for self study

4.11 References

## **4.0 OBJECTIVES**

On reading this unit, you will be in a position to explain

- ❖ Meaning of search strategy,
- ❖ Aims and principles of search strategy
- ❖ Types of search strategies
- ❖ Work involved in the preparation and how to do make a search
- ❖ What to search etc in the search strategy

## **4.1 INTRODUCTION**

In the last two decades bibliographic search systems have become extremely sophisticated. The number of databases available, their size and the differences in indexing systems have become considerably simpler and more comprehensive today when compared to their pioneer services in the 1970s. There were very few databases and their size was relatively small. The users were having very limited options. Users need to enter strings of keywords, or sets of single keywords, and combined them with Boolean operators. The ability to search one database and execute it in another database was unknown. Users had no option but to re-enter the search in each database to be searched.

Today, the scenario of database has completely changed with regard to the number of database available, their size and the search systems they offer. The search system software capabilities have grown to facilitate efficient database selection and searching with a minimum effort, say by a click of the mouse. The same search can be carried in a number of databases without re-entry of the search options. Users can even store their searches offline for further refinement for later execution. These developments cut down on the time a user must spend online to retrieve the desired information. However, there is one thing that the users cannot escape the database searching systems – building the search strategy.

## **4.2 AIMS OF SEARCH STRATEGY**

The four aims of search strategy are:

1. To match the desired number of relevant records.

2. To avoid matching irrelevant records
3. To avoid set sizes which are far too large
4. To avoid set sizes that are far too small or even empty.

### **4.3 BASIC PRINCIPLES OF SEARCH STRATEGY FORMULATION**

Search strategy encompasses all activities involved with search a database right from the reference interview with the user to the verification of final output. The knowledge of databases and their search strategy methods are essential for a librarian as well as end users.

One of the essential pre-requisites for database searching, either it is online or CD-ROM, is formulating an effective search strategy. According to Rayn E. Hoover (1982), the basic principles of an effective search strategy formulation involve

- Interview the request
- Conceptualise the search topic
- Use database vocabulary aids
- Interact with the systems
- User system capabilities

#### **4.3.1 Interview the Requester**

Interviewing the request will help to understand what a he or she wants from a search. It is like a reference interview. The narrative statement of the search topic given by the requester could be converted into a list of keywords/descriptors or search terms. If the requester is familiar with the online systems, his suggestions are useful in search strategy formulation. It is better to conduct the searches in the presence of the requester so that the search strategy could be modified suitably.

#### **4.3.2. Conceptualise the Search Topic**

Conceptualising the search topic involves analysis search request into its component parts. Some of the useful tools for conceptualizations are Venn diagrams and Boolean logic. Undesirable facts can be eliminated through the use of NOT operator of Boolean logic. It is advisable to perform the search from specific to general. Start the search with the most specific concept/aspect, if too many postings result, limit the search by combining with other concepts and if no hits, broaden the search with other concepts.

#### **4.3.3. Use Database Vocabulary Aids**

Many databases use a controlled vocabulary and/or a scheme of classification codes. Using the database vocabulary aids to identify appropriate terminology, the searcher can save time and labour. Database have built-in functions – EXPAND, NEIGHBOUR, ROOT, EXPLODE and TREE to aid the searcher.

Classification Codes are available in some of the databases and they are variously described as concept codes, subject codes, category codes, etc.,. These codes help to search broad concepts. For example, CA SEARCH uses Registry Number for chemical compounds, and BIOSIS PREVIEWS uses weighted concept codes and biosystematic codes.

#### **4.3.4. Interact with the System**

With the advances in technology, searching the online database in interactive mode and getting the results/pointout online have become simpler and cost-effective. Users get familiar with the strengths and weaknesses of the systems as well as learn tricks that the vendors do not reveal in their manuals and training sessions. It could be learnt only by interacting with the systems. Users will have various options to browse, broaden the searches, modify, restructure, regroup and resume, etc. in interactive mode.

#### **4.3.5 Use System Capabilities**

The online systems are becoming sophisticated by increasing their speed, accuracy and efficiency to access and use their database effectively. The new features allow the user to click the options and thus, the users need not type long entries at his terminal. Typographical errors could be detected by EXPAND and NEIGHBOR functions available in database indexes. The system capabilities also allow searching uncontrolled text fields such as titles and abstracts by using proximity, relational or “full-text” operators and truncation. Many databases also use online thesauri to aid the searching. Search saving and storing features on all of the CD-ROM and online systems have become efficient and powerful. Thus, the system capabilities have increased enormously over the years and only the searchers need to develop their own strategic logic to get efficient retrieval from the database.

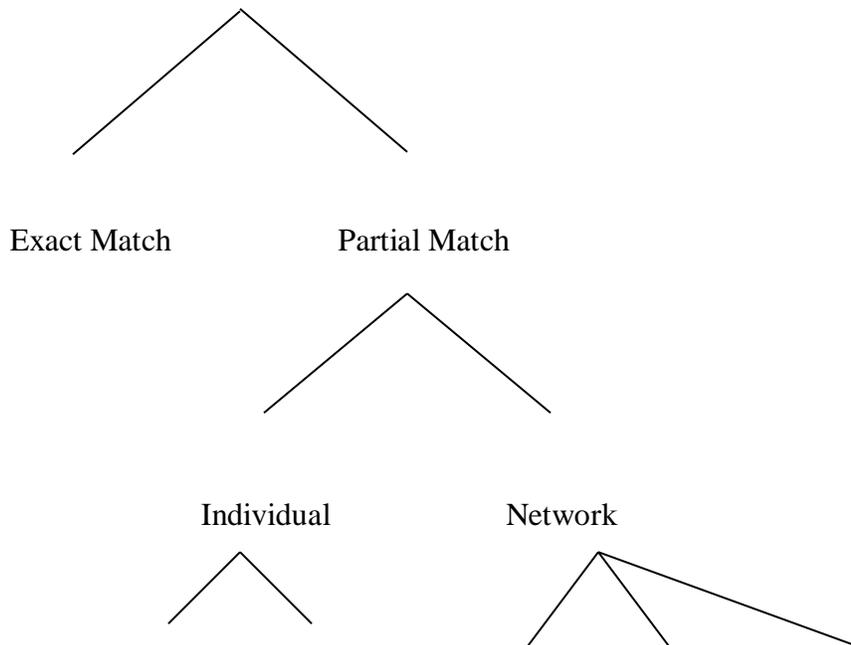
Generally the search or retrieval of information from the information retrieval system is through a query processing system. The information stored in the system is indexed using some indexing technique using key words. The processing system

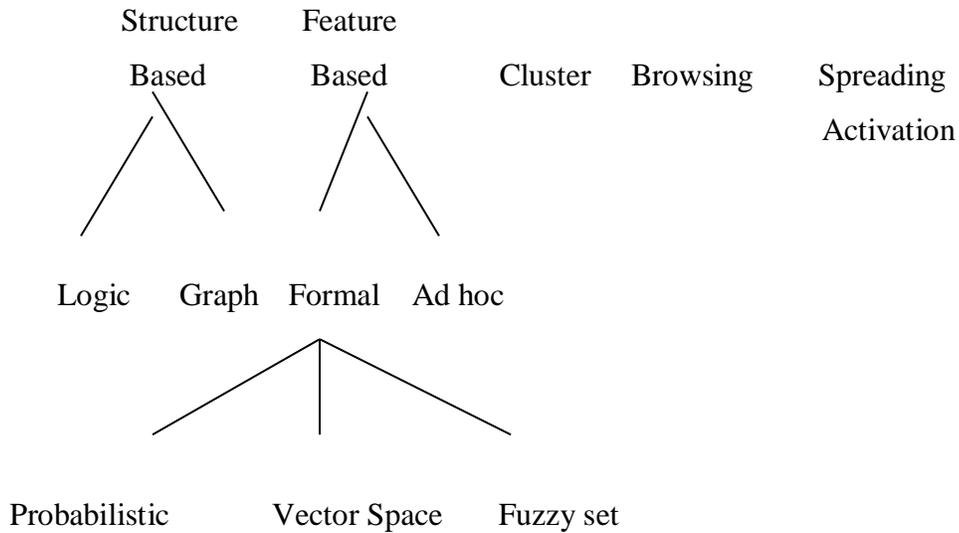
matches the key words of the query language with that of the key words under which the information items have been indexed. The matching results into response output which may be the answer to the user in responses to his request or search for information.

According to Allen Kent, any information retrieval system entails a series of processes or steps, which are as follows.

- i. Analysis involving perusal of the record and the selection of point of view (or analysis)
- ii. Terminology and subject heading control involving establishment of some arbitrary relationship among, 'analytic' in the system.
- iii. Recording the results of analysis on a searchable medium
- iv. Storage of records or source documents, involving the physical placement of the record in some locations.
- v. Questions analysis and development of search strategy involving the expression of a question or a problem.
- vi. Conducting of search involving the manipulation or operation of the search mechanism in order to identify records from the file.
- vii. Delivery of results of search involving physical removal or copying of a record from files.

#### Retrieval Techniques





**1. State the aims and principles of Search strategy**

**Note:**

- i). Write your answer in the space given below.**
- ii). Check your answer with the answers given at the end of this Unit.**

.....

.....

.....

**4.4 TYPES OF SEARCH STRATEGIES**

Search strategies can be divided into Initial Strategies and Reformation Strategies.

**Initial Strategies** are used in formulating the initial search requirements to submit to the online catalogue.

**Reformulation Strategies** are strategies for formulating subsequent search request to improve the search result after reviewing the result of the initial search.

Reformulation Strategies can be divided into: Broadening and Narrowing Strategies. Broadening Strategies are used for increasing the number of relevant records, while Narrowing Strategies are for decreasing the number of unwanted records retrieved.

Many Information Scientists view that Search strategy is very much an art rather than an exact science.

## **2. What are the types of Search strategy**

**Note:**

**i). Write your answer in the space given below.**

**ii). Check your answer with the answers given at the end of this Unit.**

### **4.5 PREPARATION OF SEARCH STRATEGY**

Preparation of a useful search strategy involves quite a number of steps. Henry et al provides a detailed checklist for search strategy, major points of which are provided below.

1. Necessary search aids:
  - System manual and search aids;
  - System and database newsletter;
  - Vocabulary control devices and classification schemes, dictionaries, glossaries, etc.
2. Information about the query:
  - Pre-search interview
  - Consultation of different reference sources;
  - Determination of exact requirement of the user in terms of query, as well as level of requirement e.g., urgency, restrictions regarding number, type, language, etc. of documents, confidentiality and other factors.
3. Whether online searching is necessary:
  - decisions regarding availability of databases on the given subject area, its coverage, cost and other related factors.
4. Choice of database governed by factors like:
  - subject coverage;
  - document coverage;
  - accessibility.
5. Decision regarding systems to use according to
  - database coverage;
  - search fields;
  - search devices;
  - performance, etc.

6. Analysis of query and selection of search terms through
  - knowledge of the systems and files;
  - use of reference tools;
7. Planning and carrying out the search through-
  - preparation of initial profile;
  - narrowing down the search by reducing Ored terms and adding ANDED terms, if necessary;
  - broadening the search by reducing the AND links, and increasing the Ored terms, if necessary;
8. Obtaining results and communicating with the user:
  - using the most suitable print format(s)
  - getting user's responses regarding the search results;
9. Recording
  - the steps followed in the search
  - the user's response regarding the search results.

### **3. How do you prepare Search strategy?**

**Note:**

- i). Write your answer in the space given below.**
- ii). Check your answer with the answers given at the end of this Unit.**

### **4.6 HOW TO SEARCH AND WHAT TO SEARCH:**

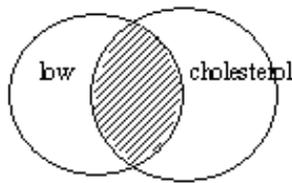
There are six steps in the process of a search namely:

- Information need
- Stated request
- Selection of a database
- Search strategy (or formulation)
- Search in database
- Screening of output

#### **4.6.1 Boolean logic**

Boolean logic is a part of a set of techniques used in mathematics for manipulating sets in a rigorous, logical fashion. It is named for the English mathematician George Boole, who developed the frame work on which it is based. Boolean logic provides three ways in which sets can be combined and online systems use all three. He devised a system of symbolic logic in which it is used three operators namely, plus (+), into (x) and hyphen(-) to combine statements in symbolic form. John Venn later expressed Boolean logic relationships diagrammatically known as Venn diagrams. The three operators of Boolean logic are the logical sum (+) , logical product (x) and logical difference (-). The three operators AND, OR, NOT provide a flexible way of combining two or more sets in order to produce a required final set. All IRS allow the users to express their queries by using these operators.

Most databases allow you to combine search terms using **Boolean logic**. Use the word **AND** to find books or articles containing **both** of your search terms. Use the word **OR** to find books or articles containing **either** of your search terms. Many databases also allow you to specify that your search terms appear **next** to or **adjacent** to each other, or **within** a certain number of words of each other.



*AND*

**low and cholesterol** retrieves records containing both *low* and *cholesterol*



*OR*

**major or minor** retrieves records containing either *major*, *minor* or both

If you do a search using the Boolean connectors **AND** as well as **OR**, be sure to use **OR** first (unless you enclose your terms with parentheses).

Example: *oil or petroleum and production*

When doing a complex search, enclose your search terms in parentheses if the database allows you to do so.

Example: (*argentina or peru*) and (*trade or commerce*) and (*canada or united states*)

### **Fields**

Each separate reference in a bibliographic database is a **record**. Each **record** consists of various **fields**: e.g., the author field, the title field, the publisher field, and the subject field. Some databases use the word *descriptors* instead of *subjects*. You can make your search more precise by specifying that you want a search term to appear in a particular field.

## **4. Write brief note on Boolean logic?**

**Note:**

- i). Write your answer in the space given below.**
- ii). Check your answer with the answers given at the end of this Unit.**

### **4.6.2 Subject headings**

By noticing the *subject headings*, or *descriptors*, assigned to a particular record, you can find ways to make your search more precise, or more comprehensive. Many databases offer a choice between searching for your term in the *subject* or in the *subject heading*. Using *subject heading* means that the term must be part of the database's official thesaurus, or list of subject headings. Using *subject* very often means that the term can appear in either the subject heading, the title, or an abstract. This is true in *WorldCat* and other *FirstSearch* databases. Also note that the terms used for subject headings vary from one database to another.

Example: Search *Three Gorges Dam* in *WorldCat* (using a **subject keyword** search).

Notice that your results do not include any publications with the words *Three Gorges Dam* in the subject heading. Try the search again using the

**subject** keywords (or **subject heading** keywords) *China* and *Dams* and *Yangtze River*.

Databases usually assign the most appropriate **specific** heading to a book or article. To find more material on your subject, try searching a **broader** or a **related** heading.

Example: To find books in JHU on the Cuban Missile Crisis, search the subject keywords *Cuban Missile Crisis* in the JHU Libraries catalog. You *will* find books that deal exclusively with the Cuban Missile Crisis, but you can expand your search by using the following related and/or broader subject headings:

*United States -- Foreign Relations -- Latin America*

*United States -- Foreign Relations -- 1961-1963*

*Kennedy, John F.*

Some databases provide a feature allowing you to search for **related subject headings**, either by clicking on the heading, or using a special function key.

#### **4.6.3 Limiting a search**

Most databases have features which allow you to limit your search (e.g., by language, date of publication, publication type, etc.). In very large databases such as *WorldCat*, limiting your search is a good idea.

Many databases allow you to do a "basic" or an "advanced" search. Choosing "advanced" will allow you the flexibility to limit your search.

#### **4.6.4 Truncation**

Most databases allow some form of truncation.

Example: *politi\** will retrieve *politics*, *political*, *politicians*, *politique*, *politische*, etc.

Different databases use different truncation symbols. See the help screens on the database you are using.

#### **4.6.5 Print resources**

Depending on your research, you may want to use print resources as well as electronic ones. The Library does have some indexes and abstracts which are not

available electronically. And, to find journal and newspaper articles written before the mid-1980s, you will almost always want to consult some print indexes.

Irrespective of the search problem online search (even a search in card catalogue / printed abstracting & indexing service) is a heuristic problem solving activity. It is difficult to develop hard and fast rules for such an activity. It will be necessary to employ heuristics if the search objectives are not met by the initial search. It is important to remember that considerable amount of training and experience is necessary to be able to carry out effective searches.

#### **4.7. Check your progress**

1. Types of Search Strategies are.....

a. Initial Strategies   b. Reformation

2. A limited set of terms that must be used to index documents in a particular system refers to

(i) Indexing

(ii) Indexing System

**(iii) Vocabulary Control**

(iv) Subject Heading.

3. The concept of 'Thesauri Facet' has been developed by

(i) Derek Austin

(ii) Ganesh Bhattacharya

(iii) Ranganathan

**(iv) Jean Atchison.**

4. The following are used as tools for vocabulary control in indexing:

1. Dictionary 2. Thesaurus 3. List of Subject Headings 4. ISBD

(a) 1 and 2 are correct

(b) 1 and 3 are correct

**(c) 2 and 3 are correct**

(d) 2 and 4 are correct

5. The merit of what type of searching is to offer the using of Boolean logic which allows limiting or expanding the search, as required?

(i) manual searching

**(ii) online searching**

(iii) literature searching

(iv) reference searching.

5. In which of the following the term “Truncation” is used

(a) Budgeting

**(b) Search Formulation**

(c) Coordination

(d) Classified Bibliography

#### **4.8.SUMMARY**

Search strategy encompasses all activities involved with search a database right from the reference interview with the user to the verification of final output. The basic principles of an effective search strategy formulation involve Interview the request; Conceptualise the search topic; Use database vocabulary aids; Interact with the systems and User system capabilities. Search strategies can be divided into Initial Strategies and Reformation Strategies. There are six steps in the process of a search namely: Information need; Stated request; Selection of a database; Search strategy (or formulation); Search in database and Screening of output.

#### **4.9 Glossary**

**Search Strategy:** The plan for how a request will be searched on the computer. It will include a series of search statements combined by Boolean operators, which will normally be planned in advance.

**Search Statement:** Instruction to the computer to find records matching the term of combination of terms entered by the searcher.

#### **4.10. Questions for self study**

1. State the aims and principles of Search strategy
2. What are the types of Search strategy?
3. How do you prepare Search strategy?
4. Write brief note on Boolean logic?

#### **4.11. REFERENCES**

Baeza-Yates, R and Ribeiro-Neto. 1999. Modern Information Retrieval. Delhi: Pearson Education (Singapore) Pte. Ltd.

Chowdhury, G. G .1994 Information Retrieval System. Calcutta: IASLIC.

Chowdhury, G. G .1999 Introduction to modern Information Retrieval System. London: Library Association Publishing.

Hartley, R. J. etal. 1993. Online searching: Principles and Practice. London: Bowker-Sour.

Lancaster, F. W. 1979. Information Retrieval Systems: Characteristics, testing and evaluation. 2<sup>nd</sup> ed. New York: John Wiley and Sons.

Walker, G and Janes, J 1999. Online Retrieval: A dialogue of theory and practice. 2<sup>nd</sup> ed. Colorado: Libraries Unlimited.



**KARNATAKA STATE OPEN UNIVERSITY  
MUKTHAGANGOTRI, MYSORE –570 006**

**Master of Library and Information Science**

**M.Lib.I.Sc - 5**

**Information Systems:  
Architecture and  
Retrieval**

**BLOCK - 2**

**BLOCK**

**2**

---

**INDEXING SYSTEMS**

---

---

**Unit -5**  
**CONTROLLED VOCABULARY TOOLS**

---

**Unit -6**  
**INDEXING LANGUAGE: CONTROLLED VS NATURAL**

---

**Unit -7**  
**SUBJECT HEADINGS LISTS, THESAURUS, THESAUROFACET**

---

**UNIT – 8**  
**INTRODUCTION TO DIGITAL LIBRARIES**

## **INSTRUCTIONAL DESIGN AND EDITORIAL COMMITTEE**

### **COURSE DESIGN**

**Prof. D. Shivalingaiah**

**Chairman**

Vice Chancellor  
Karnataka State Open University  
Mukthagangotri, Mysuru-570006

**Prof. M. Mahadevi**

**Convener**

Dean (Academic)  
Karnataka State Open University  
Mukthagangotri, Mysuru-570006

### **COURSE COORDINATOR**

**Shilpa Rani N R**

Chairperson

Department of Studies in Library and Information Science  
Karnataka State Open University, Mukthagangotri, Mysuru-570006

### **COURSE EDITORS**

**Prof. M A Gopinath**

Professor (Retd.) in LISc  
DRTC, ISI Building, Mysore Road,  
Bangalore

**Prof. A Y Asuudi**

Professor (Retd.) in LISc  
Bangalore University  
Bangalore

**Dr. N. S Harinarayana**

Senior Lecturer  
Dept. of Library & Information Science  
University of Mysore, Mysore -06

**Prof. V. G. Talwar**

Professor in LISc  
Dept. of Library & Information Science  
University of Mysore, Mysore -06

### **COURSE WRITER**

**Prof. N BPangannaya**

Retd. Professor  
University of Mysore

### **BLOCK EDITOR**

**Prof. N BPangannaya**

Retd. Professor  
University of Mysore

**Dr. P G Tadasad**

Lecturer (Sr Scale)  
Dept. of Library & information science  
Gulbarga University, Gulbarga

**Prof. V G Talawar**

Professor  
Dept. of Library & information science  
University of Mysore, Mysore -06

### **PUBLISHER**

**Registrar**

Karnataka State Open University  
Mukthagangotri, Mysuru-570006

Developed by Academic Section KSOU, Mysore

**Copy Right: KARNATAKA STATE OPEN UNIVERSITY, 2017**

© All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from Karnataka State Open University.

This courseware is printed and published by The Registrar, Karnataka State Open University, Mysuru for limited use only. No individual or collaborative institution can use / print / distribute in any form without the written permission from KSOU. For user rights of this content and for

other queries contact The Planning and Development Officer, KSOU, Mysuru 570 006. Digital delivery of this courseware is also available for those who opt. For more details visit [www.ksoustudymaterial.com](http://www.ksoustudymaterial.com) or [www.ksoumysore.edu.in/digitalcontent](http://www.ksoumysore.edu.in/digitalcontent)

---

## **M.Lib.I.Sc - 5: Information Systems: Architecture and Retrieval**

### **Block – 2 : Indexing Systems**

---

#### **Block Introduction**

The objectives of the Course 5 – Information systems: Architecture and Retrieval is to help students develop familiarity with concepts, principles, and components of IRS, various tools, retrieval models, different method and techniques. In Block one of this course you have already been introduced to the concept of IRS. In this Block, you will be explained the concept of Indexing Systems.

In Unit 5 the need for vocabulary control, how vocabulary control is achieved, purpose of controlled vocabularies, impact of vocabulary control on information retrieval various vocabulary control tools, principles of vocabulary control tools, structures of Controlled Vocabularies viz., List, Synonym ring, Taxonomy and Thesaurus, semantic relationships used in the tools and the general considerations to be considered for displaying are discussed

Unit 6 describes the functions/ uses of Index, traces the history of indexing, lists types of indexes, and discusses the role of indexing in information retrieval. This unit also traces the evolution of indexing systems from Cutter to Austin and describes the process of indexing. Different types of indexing systems - Pre-Coordinate, Post-Coordinate and Titlebased Indexing Systems are explained in detail. Types of indexing languages (Natural language, Free and Controlled indexing language) and characteristics of indexing language are also discussed. Important indexing models like Chain Indexing, Preserved Context Indexing System (PRECIS), Postulate-based Permuted Subject Index (POPSI) and Keyword in Context Indexing System are described in detail.

In Unit 7 salient features of two major subject headings lists namely Library of Congress Subject Headings (LCSH) and Sears Subject Headings are discussed. The concept of thesaurus, its construction and two important thesauri, MeSH and INSPEC are also described in detail. the concept of thesaurus facet including BSI ROOT Thesaurus is also explained. At the end the students are introduced to the concept of classarus.

Unit 8 in this Block presents the evolution, definition and characteristics, purpose and features of digital libraries and how they are different from virtual libraries and their advantages.

Librarians play the role of connecting information and people together. In this new role libraries can adopt all the internet, technical, computing and architectural knowledge required. **J Veen (2001)** defines the role of information architect as: 1). the individual who organizes the patterns inherent in data, making the complex clear; 2) a person who creates the structure or map of information, which allows others to find their personal paths to knowledge”.

**Prof. V G Talawar**

---

---

## **MLISc – 5**

### **Information Systems: Architecture and Retrieval**

---

---

## **Block – 2**

### **Indexing Systems**

---

---

## **Unit – 5**

### **Vocabulary Control Tools**

---

#### **Structure:**

5.0 Objectives

5.1 Vocabulary Control

5.11 Need for Vocabulary Control

5.12 How vocabulary control is achieved

5.13 Purpose of Controlled Vocabularies

5.14 Impact of Vocabulary Control on Information Retrieval

5.2 Vocabulary Control Tools

5.3 Principles of Vocabulary Control Tools

5.31 Ambiguity

5.32 Synonymy

5.33 Semantic Relationships:

5.34 Using Warrant to Select Terms

5.341 Literary Warrant:

5.342 Organizational Warrant

5.343 User Warrant

## 5.4 Structures of Controlled Vocabularies

### 5.41 List

### 5.42 Synonym ring

### 5.43 Taxonomy

### 5.44 Thesaurus

## 5.5 Semantic Relationships used in Controlled Vocabularies

### 5.51 Semantic Linking

### 5.52 Indicating Relationships Among Terms

## 5.6 General Considerations

### 5.61 Elements to Address

### 5.62 User Categories

## 5.7 Check your progress

## 5.8 Summary

## 5.9 Glossary

## 5.10. Questions for self study

## 5.11 References

## **5.0 OBJECTIVES:**

By studying this unit you will be able to

- ❖ Understand the concept of vocabulary control
- ❖ Justify the need for and purpose of vocabulary control
- ❖ Analyze the impact on information retrieval
- ❖ Know different types of vocabulary control tools
- ❖ Identify the principles and structure of such tools
- ❖ Know the semantic relationships used in the tools and
- ❖ The general considerations to be considered for displaying

## **5.1 VOCABULARY CONTROL:**

You have studied in Block I of this course that while searching an information retrieval system one will try to achieve efficiency in two ways. Retrieve as many relevant items as possible, which is called the recall of the search and also want only relevant

items in answer. This is called the precision of search. This process has limitations, caused by certain characteristics of language. Any language is flexible enough to incorporate new terms and syntax to represent correct association of terms and specify any subject. But specifying a subject is not enough for organizing an index file. For a natural language this flexibility is considered as its richness. But this richness of natural language is a hindrance for using it in organizing index files. For example, language contains synonyms: that is different words that have the same meaning. Language also contains homographs: words that are spelled the same but which have different meanings. Also the same concept may be identified by different names. The terminology may differ in different countries. Separate terms might have been used in popular and descriptive languages.

When different users have different names for the same thing, the resulting confusion reduces effective communication. Then why can't everyone just use the same vocabulary? This problem is difficult because everyone comes to their work with different educational backgrounds and experience, and they may have different names for similar subjects or concepts. Having a standard controlled vocabulary and a standard set of practice descriptors, organized into a standard arrangement, takes a lot of discipline. Sometimes people need training in the accurate spelling of terms important in their workplace. This can be tackled by making a concept controlled. A vocabulary is said to be controlled if it consists of a restricted subset of possible terms. Such a subset, in that it contains only those terms “authorized” for use, is sometimes called an authority list. In addition to terminological restriction, most vocabulary control tools articulate semantic relationships between terms in the vocabulary, the most common of these being the inclusion of hierarchical relationship.

Designing the controlled vocabularies using vocabulary control tools can do this. “Vocabulary control is the sine qua non of information organization,” says Elaine Svenonius. Vocabulary control is used to improve the effectiveness of information storage and retrieval systems, Web navigation systems, and other environments that seek to both identify and locate desired content via some sort of description using language.

The primary purpose of vocabulary control is to achieve consistency in the description of content objects and to facilitate retrieval.

**5.11 Need for Vocabulary Control:** The need for vocabulary control arises from two basic features of natural language, namely

- Two or more words or terms can be used to represent a single concept

Example:       Salinity/saltiness  
                  VHF/Very High Frequency

- Two or more words that have the same spelling can represent different concepts

Example:       Mercury (planet)  
                  Mercury (metal)  
                  Mercury (automobile)  
                  Mercury (mythical being)

## **5.12 How vocabulary control is achieved**

Vocabulary control is achieved by three principal methods:

- Defining the scope, or meaning, of terms
- Using the equivalence relationship to link synonymous and nearly synonymous terms;  
and
- Distinguishing among homographs

The guidelines for constructing the controlled vocabularies are as follows

- selecting the terms
- formulating the terms
- establishing relationships among terms, and
- presenting the information effectively in printed, online, and web navigation sites.

### 5.13 Purpose of Controlled Vocabularies

The purpose of controlled vocabularies is to provide a means for organizing information. Through the process of assigning terms selected from controlled vocabularies to describe documents and other types of content objects, the materials are organized according to the various elements that have been chosen to describe them.

Controlled vocabularies serve five purposes:

1. **Translation:** Provide a means for converting the natural language of authors, indexers, and users into a vocabulary that can be used for indexing and retrieval.
2. **Consistency:** Promote uniformity in term format and in the assignment of terms.
3. **Indication of relationships:** Indicate semantic relationships among terms.
4. **Label and browse:** Provide consistent and clear hierarchies in a navigation system to help users locate desired content objects.
5. **Retrieval:** Serve as a searching aid in locating content objects.

### 5.14 Impact of Vocabulary Control on Information Retrieval

Information retrieval effectiveness is traditionally measured by two parameters: recall and precision. Controlled vocabulary design can have a positive impact on both of these measures.

**5.141 Recall:** Recall can be improved through such controlled vocabulary methods as:

- Preferred terms and equivalence relationships for synonym control
- Preferred term form
- Associative (related term) relationships
- Classified and hierarchical relationships
- Postcoordination
- Concept mapping / clustering

**5.142 Precision:** Precision can be improved through such controlled vocabulary methods as:

- Parenthetical qualifiers to control ambiguity
- Broader and narrower term hierarchical relationships
- Compound terms
- Precoordination

## **5.2 VOCABULARY CONTROL TOOLS:**

Controlled vocabulary tools mandates the uses of predefined, authorised terms that have being preselected by the designer of the controlled vocabulary as opposed to natural language vocabularies where there is no restriction on the vocabulary that can be used. Controlled vocabulary is a carefully selected list of words and phrases, which are used to tag units of information (document or work) so that they may be more easily retrieved by a search. Controlled vocabularies solve the problems of homographs, synonyms and polysemes by ensuring that each concept is described using only one authorized term and each authorised term in the controlled vocabulary describes only one concept. In short, controlled vocabularies reduces ambiguity inherent in normal human languages where the same concept can be given different names and ensures consistency. There are many vocabulary control tools that include spell-checkers, proper noun, and date identification, synonym list, authority lists, subject headings, thesauri, indexing schemes and taxonomies (You will be studying them in the next two units in detail).

**5.21 Authority lists:** Authority lists also called authority files are lists of terms that can be used to control the terms or variants that are used in their collections documentation. For example, one may use an authority list during data entry, in order to ensure that the name is spelled consistently, or to ensure that a certain version is consistently used. There may or may not be a "preferred" variant of the term, but all variants are linked in the authority list so that the term can be found.

**5.22 Thesauri:** Thesauri usually provide synonyms, broader and narrower terms, and "preferred" terms for concepts. Some thesauri also include scope notes to advise cataloguers of the precise meaning and usage of particular concepts found in the thesaurus. Thesauri may be used in a similar way as a search assistant - people can use the thesaurus to find the most effective search terminology for a particular concept.

**5.23 Subject heading system:** In subject heading system that uses controlled vocabulary, authorised terms (subject headings in this case) have to be chosen to handle choices between variant spellings of the same concept (American versus British), choice among scientific and popular terms (Cockroaches versus *Periplaneta americana*), choices between synonyms [automobile versus cars] among other difficult issues. Choice of authorised terms are based on the principles of user warrant (what terms users are likely to use), Literacy warrant (what terms are generally used in the literature and documents), structural warrant (terms chosen by considering the structure, scope of the controlled vocabulary). Controlled vocabularies also typically handle the problem of homographs, with qualifiers. For example, the term "pool" has to be qualified to refer to either Swimming pool, or the game pool to ensure that each authorised term or heading refers to only one concept.

Historically Subject headings was designed to describe books in library catalogs by catalogers while thesauri were used by indexers to apply index terms to documents and articles. Subject headings tend to be broader in scope describing whole books, while thesauri tended to be more specialised covering very specific disciplines. Also because of the card catalog system, subject headings tends to have terms that are in indirect order (though with the rise of automated systems this is being removed), while thesauri terms are always in direct order. Subject headings also tend to use more pre-co-ordination of terms such that the designer of the controlled vocabulary will combine various concepts together to form one authorised subject heading. (e.g Children and terrorism) while thesauri tend to use singular direct terms. Lastly thesauri lists not only equivalence terms but also narrower, broader terms, related terms among various authorised and non-authorised terms, while historically most subject headings did not. Well known subject

heading systems are Library of Congress Subject Heading, MESH, Sears List of Subject Headings and SHE. Well known thesauri are Art and Architecture Thesaurus, ERIC Thesaurus etc. Controlled vocabularies tagged to documents are metadata.

**5.24 Classification Schemes:** Classification schemes differ from Subject Headings in that they are designed primarily to provide a means of identifying an appropriate location on shelves for a book, e.g. books with the same 'main' subject area grouped together on shelves. However the place of any particular book in any particular library may be different, as different libraries will try to organize their shelves to meet the needs of their users. Classification, involves the development and use of a scheme for the systematic organization of knowledge. There are three approaches to classification: enumerative, hierarchical, and analytico-synthetic. Enumerative classification attempts to assign headings for every subject and alphabetically enumerates them. Hierarchical classification uses a more philosophical approach based on the inherent organization of the subject being classified, and establishes logical rules for dividing topics into classes, divisions, and subdivisions. Analytico-synthetic classification assigns terms to individual concepts and provides rules for the local cataloger to use in constructing headings for composite subjects. Though these are basically used for retrieval purpose, at the semantic level these can also act as vocabulary control tools. Library of Congress Classification, Dewey Decimal Classification, Universal Decimal Classification and Colon Classification are few of the important schemes.

### **5.3 PRINCIPLES OF VOCABULARY CONTROL TOOLS:**

There are four important principles of vocabulary control that guide their design and development.

These are:

- eliminating ambiguity
- controlling synonyms
- establishing relationships among terms where appropriate
- testing and validation of terms

A major goal of vocabulary control is to ensure that each distinct concept refers to a unique linguistic form. These types of linguistic relationships should be controlled or regularized so that information or content that is provided to a user is not spread across the system under multiple access points, but is gathered together in one place. Eliminating ambiguity and compensating for synonymy through vocabulary control assures that each term has only one meaning and that only one term may be used to represent a given concept or entity.

**5.31 Ambiguity:** Ambiguity occurs in natural language when a word or phrase (a homograph or polyseme) has more than one meaning. Figure 2 provides an example and shows how a single word may be used to represent multiple, very different concepts.

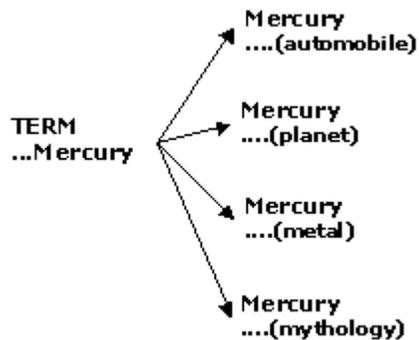


Figure 1: Ambiguity caused by homographs and polysemes

A controlled vocabulary must compensate for the problems caused by ambiguity by ensuring that each term has one and only one meaning.

**5.32 Synonymy:** A different problem occurs when a concept can be represented by two or more synonymous or nearly synonymous words or phrases. This is called synonymy. This means that desired content may be scattered around an information space or database because it can be described by different but equivalent terminology. Figure 3 illustrates this case:

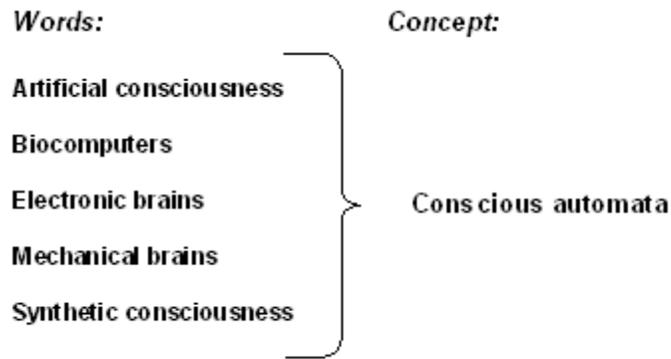


Figure 2: Information scatter caused by synonyms

A controlled vocabulary must compensate for the problems caused by synonymy by ensuring that each concept is represented by a single preferred term. The vocabulary should list the other synonyms and variants as non-preferred terms with USE references to the preferred term.

**5.33 Semantic Relationships:** Various types of semantic relationships may be identified among the terms in a controlled vocabulary. These include equality relationships, hierarchical relationships and associative relationships, which may be defined as required for a particular application.

**5.34 Using Warrant to Select Terms:** The process of selecting terms for inclusion in controlled vocabularies involves consulting various sources of words and phrases as well as criteria based on:

- the natural language used to describe content objects (literary warrant),
- the language of users (user warrant), and
- the needs and priorities of the organization (organizational warrant).

**5.341 Literary Warrant:** Assessing literary warrant involves consulting reference sources such as dictionaries or textbooks as well as existing vocabularies. The word or phrases chosen should match as closely as possible the prevailing descriptions for the

concept in the literature. Terminology should be drawn from canonical examples of writing in our content management community.

**5.342 Organizational Warrant:** Determining organization warrant requires identifying the form or forms of terms that are preferred by the organization or organizations that will use the controlled vocabulary. There may be intrinsic natural structures that should be reflected in the choice of terms.

**5.343 User Warrant:** This is the origin of today's mantra of "user-centered design." We need to study the vocabulary in use today by our community. Creating lists of potential terms to enhance completeness of the vocabulary.

- Organizing candidate terms into broad categories to determine what categories users prefer and what they should be called.
- Placing candidate terms into a tentative set of broad categories to validate categories that have been created.
- Reviewing drafts of the vocabulary to add missing terms, delete terms that are incorrect or obsolete, create more useful term forms, and identify and correct missing and/or incorrect relationships among terms.

#### **5.4 STRUCTURES OF CONTROLLED VOCABULARIES:**

Controlled vocabularies are structured to enable displaying the different types of relationships among the terms they contain. There are four different types of controlled vocabularies, determined by their increasingly complex structure. These are:

- List
- Synonym ring
- Taxonomy
- Thesaurus

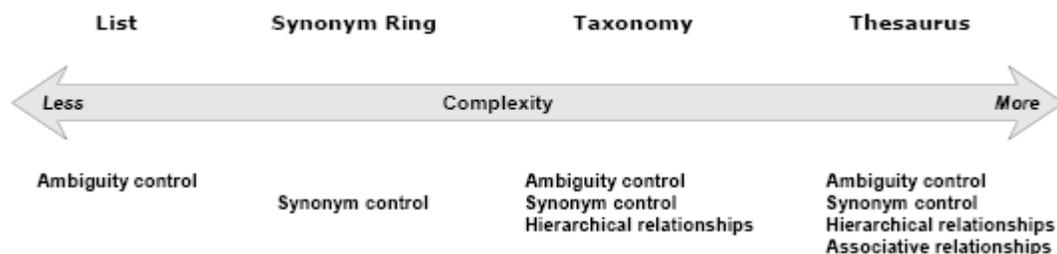


Figure 3: Increasing structural complexity among controlled vocabularies

This figure shows the increasingly complex structure of controlled vocabularies dictated by the requirements of the types of relationships each must accommodate. It also shows that the more complex vocabularies (taxonomies, thesauri) include the simpler structures (lists, synonym rings). For example, a Thesaurus includes explicit devices for controlling synonyms, arranging hierarchies, and creating associative relationships while a List is a simple set of terms containing no relationships of any kind.

**5.41 List:** A list (also sometimes called a pick list) is a limited set of terms arranged as a simple alphabetical list or in some other logically evident way. Lists are used to describe aspects of content objects or entities that have a limited number of possibilities. Examples include geography (e.g., country, state, city), language (e.g., English, French, Swedish), or format (e.g., text, image, sound).

**Example 1: Simple alphabetical list**

Alabama  
Alaska  
Arkansas  
California  
Connecticut  
Delaware

**Example 2: Simple logical list**

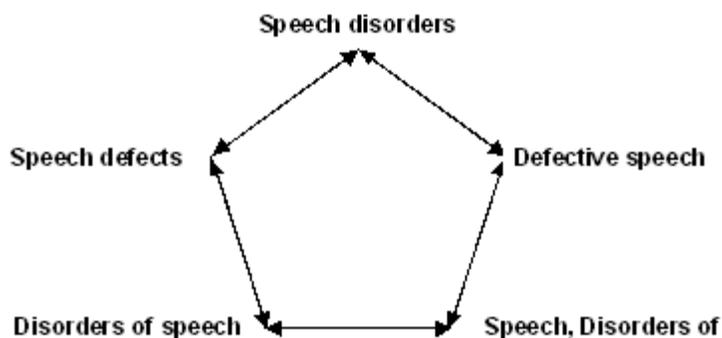
Mercury  
Venus  
Earth  
Mars  
Jupiter  
Saturn  
Uranus  
Neptune  
Pluto

**5.42 Synonym ring:** While a synonym ring is considered to be a type of controlled vocabulary, it plays a somewhat different role than the other types covered by this Standard. Synonym rings cannot be used during the indexing process. Rather, they are

used only during retrieval. Use of synonym rings ensures that a concept that can be described by multiple synonymous or quasi-synonymous terms will be retrieved if any one of the terms is used in a search.

A synonym ring, therefore, is a set of terms that are considered equivalent for the purposes of retrieval. Synonym rings usually occur as sets of flat lists. A synonym ring allows users to access all content objects or database entries containing any one of the terms. Synonym rings are generally used in the interface in an electronic information system, and provide access to content that is represented in natural, uncontrolled language.

Example 3: Synonym ring for speech disorders



**5.43 Taxonomy:** A taxonomy is a controlled vocabulary consisting of preferred terms, all of which are connected in a hierarchy or polyhierarchy.

Example 4: Taxonomy hierarchy



**5.44 Thesaurus:** A thesaurus is a controlled vocabulary arranged in a known order and structured so that the various relationships among terms are displayed clearly and identified by standardized relationship indicators. Relationship indicators are usually employed reciprocally.

Example 5: Print thesaurus entry:

## ABSORPTION

The retention and conversion into another form of energy of rays, waves, or particles by a substance.

UF ABSORPTIVE PROPERTIES

BT **SORPTION**

NT **BIOLOGICAL ABSORPTION**

**RESONANCE ABSORPTION**

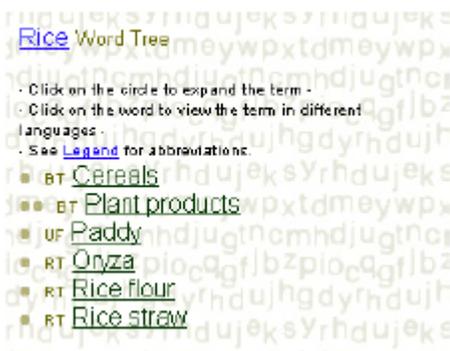
**TWO PHOTON ABSORPTION**

**X RAY ABSORPTION ANALYSIS**

Source: DTIC Thesaurus

Example 6: Online thesaurus entry:

### Example 6: Online thesaurus entry



Source: AGROVOC

## 5. 5 SEMANTIC RELATIONSHIPS USED IN CONTROLLED VOCABULARIES:

There are three types of relationships used in controlled vocabularies:

- a) Equivalency
- b) Hierarchy
- c) Association

**5.51 Semantic Linking:** The relationships among terms in a controlled vocabulary are indicated by semantic linking. Semantic linking encompasses various techniques and conventions for indicating the relationships among terms.

Table 1 The basic types of relationships and provides some simple examples.

Relationship Type	Example
<b>Equivalency</b>	
Synonymy	UN / United Nations
Lexical variants	pediatrics / paediatrics
Near synonymy	sea water / salt water smoothness / roughness
<b>Hierarchy</b>	
Generic or IsA	birds / parrots
Instance or IsA	sea / Mediterranean Sea
Whole / Part	brain / brain stem
<b>Associative</b>	
Cause / Effect	accident / injury
Process / Agent	velocity measurement / speedometer
Process / Counter-agent	fire / flame retardant
Action / Product	writing / publication
Action / Property	communication / communication skills
Action / Target	teaching / student
Concept or Object / Property	steel alloy / corrosion resistance
Concept or Object/ Origins	water / well
Concept or Object / Measurement Unit or Mechanism	chronometer / minute
Raw material / Product	grapes / wine
Discipline or Field / Object or Practitioner	neonatology / infant

**5.52 Indicating Relationships Among Terms:** A basic property of relationships in controlled vocabularies is that they are reciprocal; that is, each relationship indicated between Term A and Term B has a corresponding relationship from Term B to Term A. This rule is observed for all types of relationships.

The conventional abbreviations for relationship indicators are used in the examples below. Additional abbreviations for specialized purposes are found in the following sections.

The relationship indicators are always paired operators. Some indicators are symmetric while others are asymmetric as illustrated below:

- Related Term (RT) is symmetric:

If **Term A RT Term B**, then **Term B RT Term A**

- Preferred Term (Equivalency) – USE and UF are asymmetric:

If **Term A** USE **Term B**, then **Term B** UF **Term A**

- Hierarchical Relationships – BT and NT:

If **Term A** BT **Term B** then **Term B** NT **Term A**.

The way in which a controlled vocabulary is presented affects the user's willingness and ability to make use of it. This section defines requirements and recommendations for effective display of controlled vocabularies. Note that in some applications, portions of a controlled vocabulary may be displayed in connection with the content objects they reference. However, this display is governed by the software used to generate the search results and is not covered by this Standard.

## 5.6 GENERAL CONSIDERATIONS

**5.61 Elements to Address:** The vocabulary developer should address the following elements of the vocabulary display:

- **Presentation:** Presentation decisions include how to represent the term relationships, whether and how to make typography distinctions, capitalization, and filing rules.
- **Type of Display:** There are a large variety of display types that can be used, from simple alphabetical listings to complex graphical displays. Often multiple structures are presented to the user.
- **Format:** Controlled vocabularies may be delivered in print or electronic formats or both. Web and hyper-linking technology allows additional display capabilities not available in print form.
- **Documentation:** All controlled vocabularies should provide user documentation that describes how to use the vocabulary.

**5.6.2 User Categories:** The design of displays for controlled vocabularies should take into account the needs of each anticipated class of user:

**a) Controlled vocabulary maintenance personnel** – These users have expertise in indexing and controlled vocabulary construction and are likely to be experts in the subject domain of the controlled vocabulary. They must have access to all views of a controlled vocabulary and complete information about each term, with the ability to edit and manipulate term records, cross-references, classification notation, and hierarchies. They require “housekeeping” displays not needed by end users of a controlled vocabulary.

**b) Indexers and expert searchers** – These users have expertise in indexing, online information retrieval, use of controlled vocabularies, or all of these. Indexers are likely to have expertise in the subject domain of the controlled vocabulary while expert searchers may or may not have such expertise. These sophisticated users require the ability to search and view cross-references, definitions, and notes for terms as well as various levels of the classification or hierarchies. Postings data are especially important for searchers. Sophisticated controlled vocabulary displays and terminology are appropriate for these users.

**c) End users** – These users are not likely to be experienced in the jargon and complexities of online information retrieval or the conventions of controlled vocabulary notation. They may have expertise in the subject field and understand its terminology. The types of displays available to expert searchers can be useful to end users as well, when designed with their needs in mind. End users can benefit from on-screen instructions in addition to any printed documentation that exist.

Controlled vocabulary developers may want to produce different versions of the vocabulary:

- a basic list of terms, references, and relationships designed for the end user or occasional searcher, and
- a more complete version designed for the indexer and the expert searcher, which may include detailed scope notes, indexing instructions, information on term history, and postings data.

### III Self Check exercise:

Compare your answers with the one given at the end

Write answers in the given space only

### 5.7. Check your progress

1. A number of devices that are related to recall and precision have been studied by

(i) Gilchrist (ii) Vickery **(iii) Lancaster** (iv) Ranganathan

2. The relationship between the terms 'Ships' and 'Boat' is

**(i) Syntactic relation** (ii) Semantic relation (iii) Both of them (iv) Neither

3. Requesting a borrower to return the book before its due date is called

(i) holds (ii) reserve **(iii) recall** (iv) lending.

4. The use of more words or symbols than are necessary to convey a meaning results in

(A) Efficiency

(B) Accuracy

**(C) Redundancy**

(D) Precision

5. An increase in the level of 'specificity' of indexing languages results in increase in

(a) Recall **b) Precision** (c) Noise (d) both recall and precision

### 5.8 Summary:

In this unit the concept of vocabulary control has been explained and the need for vocabulary control is justified. Further the purpose of vocabulary control is discussed. The impact of controlled vocabulary on information retrieval is also explained. An attempt has also been made to introduce different types of vocabulary control tools and provide principles and structures of such tools. The semantic relationships used in the controlled vocabularies and the general considerations to be made for displaying have also been discussed.

### 5.9 Questions for self study

1. Justify the need for vocabulary control?
2. Explain the purpose of vocabulary control?
3. Discuss the impact of vocabulary control on information retrieval?
4. Mention the different types of vocabulary control tools?
5. Explain the principles of vocabulary control tools?
6. What do you mean by user warrant?
7. Mention the different structural approaches of controlled vocabularies??
8. Explain the semantic relationships used in the controlled vocabularies?
9. What are the general criteria to be considered while designing the format of the display?

## **5.9 REFERENCES**

1. ANSI/NISO Z39.19-2005. Guidelines for the construction, format and management of monolingual controlled vocabularies
2. Chakraborty, A R and Chkraborthy, Bhubaneswar. Indexing: Principles, Process and Product. Calcutta: The World Press, 1984
3. Fast, K.; Leise, F. and Steckel. M. (2002). What is a Controlled Vocabulary? [http://web.archive.org/web/20030811115443/http://www.boxesandarrows.com/archives/what\\_is\\_a\\_controlled\\_vocabulary.php](http://web.archive.org/web/20030811115443/http://www.boxesandarrows.com/archives/what_is_a_controlled_vocabulary.php) [Accessed on 24.02.2007]
4. Fosket A C. Subject approach to information, Ed 5. London: Library Association, 1996
5. Ghosh, S K and Satpathi, J N. Subject Indexing Systems: Concepts, Methods and Techniques. Calcutta: IASLIC, 1998
6. Hutchins, W. J. Languages of indexing and classification. A linguistic study of structures and functions. London: Peter Peregrinus, 1975.
7. Lancaster, F. W. Vocabulary Control for Information Retrieval. 2nd ed. Info Resources Press, 1986.

8. Lancaster, F. W. *Indexing and abstracting in theory and practice*. London: Facet Publishing, 2003.
9. Vickery, B. C. *Information Systems*. London: Butterworth, 1973.
10. "[http://en.wikipedia.org/wiki/Controlled\\_vocabulary](http://en.wikipedia.org/wiki/Controlled_vocabulary)
11. <http://www.controlledvocabulary.com> [Accessed on 23.02.2007]
12. <http://www.slis.kent.edu/~mzeng/Z3919/> [Accessed on 23.02.2007]

---

## **UNIT 6: Indexing Languages: Natural Vs Controlled**

---

### **Structure:**

#### 6.0 Objectives

#### 6.1 Index and Indexing

6.1.1 Functions/ Uses of Index:

6.1.2 History of Indexing:

6.1.3 Types of Indexes:

6.1.4 Role of Indexing in Information Retrieval:

6.1.5 Evolution of Indexing Systems:

6.1.6 Process of Indexing:

#### 6.2 Indexing Systems

6.2.1 Pre-Coordinate Indexing system:

6.2.2 Post-Coordinate Indexing Systems:

6.2.3 Titlebased Indexing Systems

#### 6.3 Indexing Language:

6.3.1 Types of indexing language

6.3.1.1 Natural language indexing language

6.3.1.2 Free indexing languages:

6.3.1.3 Controlled indexing language

#### 6.4 Characteristics of Indexing Language:

#### 6.5 Indexing models

6.5.1 Chain Indexing:

6.5.1.1 Steps in Chain Procedure:

6.5.1.2 Comments:

6.5.2 Preserved Context Indexing System (PRECIS):

6.5.2.1 Theoretical Framework:

6.5.2.2 Format

6.5.2.3 Syntax, Semantics, Vocabulary and Cross references:

6.5.2.4 Comments:

### 6.5.3 Postulate-based Permuted Subject Index (POPSI):

#### 6.5.3.1 Theoretical Foundations:

#### 6.5.3.2 Syntax, Semantics and Sequencing of terms:

#### 6.5.3.3 Steps in POPSI:

#### 6.5.3.4 POPSI-Specific:

#### 6.5.3.5 Comments:

### 6.5.4 Keyword in Context Indexing System

#### 6.5.4.1 Structure

#### 6.5.4.2 Format

#### 6.5.4.3 Variations of Keyword Index

#### 6.5.4.4 comments

### 6.6 Check your progress

### 6.7. Summary

### 6.8. Glossary

### 6.9. Questions for self study

### 6.10. References

---

## **6.0 OBJECTIVES:**

After studying this unit, you should be able to:

- understand the basic concepts like ‘Index and indexing’
- trace the evolution of indexing system
- know the process of indexing
- differentiate between different kinds of indexing systems – Pre-coordinate and Post-coordinate
- identify the types and characteristics of indexing languages
- describe the salient features of important indexing models viz., Chain indexing, PRECIS, POPSI and KWIC

## **6.1 INDEX AND INDEXING**

To retrieve desired information from any large collection of documents, we require two things of that collection. First the collection must be in an order and second we must have a means of searching and matching; there should be some way of recognizing whether or not any given document contains the information we want. Reading the complete document is one approach, but it involves time factor and it still does not guarantee the searcher the required Information. To resolve both the time and the recognition problems, we use an Index, as a tool for locating desired Information.

The word Index is derived from the Latin word 'Indicare' which means to indicate or point out. According to Encyclopedias of Library of Information Science an index is "a systematic guide to an item contained in or concepts derived from a collection." The items or derived concepts are represented by entries arranged in a known or stated order, such as alphabetical, chronological or numerical. An index gives the thought content of both documentary and non-documentary sources. Index is prepared for the user to help him in finding information more quickly and easily. An index is a list of pointers. More specifically, it is an ordered list of terms or phrases (denoting authors, titles, or concepts) paired with pointers to content within a set of one or more documents where the terms or phrases are the focus of the content.

The preparation of a series of entries for inclusion in a subject catalogue or printed index is known as 'indexing'. Kaiser defined indexing as "the process by which information is made accessible". Indexing is a technique providing service operations and an index or a subject catalogue is a tool. Indexing is the process of organizing the entries in a file in a predetermined order and supplementing the entries with all sorts of cross references that is organizing a searchable file is the sole purpose of indexing. It is very much akin to the principles of cataloging and classification. But unlike cataloging and classification, it focuses less on describing the "aboutness" of documents and more on the extraction of terms and page numbers from texts and arranging these items into an organized structure. This structure is used to refer the reader to particular section of the text. Hence, indexing provides "searchability" to a text. One can search 'Encyclopedia Britannica' because it has an index.

### **6.11 Functions/ Uses of Index:**

The function of an Index is to provide users with an efficient and systematic means for locating documents or parts of documents that may address information needs or requests. An index should therefore

1. Identify and locate potentially relevant Information in the document or collection being indexed.
2. Discriminate between information on a topic and passing mention of a topic.
3. Analyze concepts treated in a document so as to produce suitable index headings based on its terminology.
4. Indicate relationships among topics.
5. Group together information on topics scattered by the arrangement of the document or collection.
6. Organize headings and their modifying subheadings into index entries.
7. Direct users seeking information under terms not chosen as index headings to headings that have been chosen, by means of 'see references'.
8. Suggest users of a topic to lookup related topics also by means of see also references.  
Ex: Cow see also Mammals.
9. Arrange entries into a systematic and helpful order.
10. Provide guide to material that the user may wish to recall or that he may not know exists; that is Indexes are used for questions of recall or discovery.
11. Provide a general view of a subject can be obtained from a subject Index.

12. Aid in solving the problem presented by the many languages in which material is now published. Indexes, in one language, serve as guides to the material and help the searcher to determine the need to consult the original article.
13. Provide solution to the problem of enormous number of documents published annually, to facilitate rapid selection of relevant material.

### **6.12 History of Indexing:**

The history of indexing can date back to 2<sup>nd</sup> Millennium BC. Mesopotamian cuneiform documents were enclosed in clay envelopes, which serve as an Index of sorts. By 12<sup>th</sup> C. people started taking interest in indexing as a tool. True Alphabetical Indexing, seems to have emerged in the 14<sup>th</sup> C. These consisted primarily of the keywords in the theses or disputations, alphabetically arranged. Mean while the first catalogue with an index turned up in the 15<sup>th</sup> C. In 1545 Conrad Gessner's Bibliotheca Universalis listed the documents under the alphabetical order of the author's forename. Later in 1548, it indexed the same documents in a subject classification order with an alphabetic subject index to classification codes. This can be considered as the genesis of all the present indexing system and Technologies. In 1856, Andrea Crestadoro made an attempt to show the importance of titles of documents in cataloging work. Later in 1959 H P Luhn of IBM developed KWIC (Keyword in Context) by using computers. From 1970s the rise of Selective Dissemination of Information (SDI) services, titles of scientific documents began to play a significant role in science communication. The title based Indexes depend upon manipulation of all keywords the title to give multiple entries one entry for each significant word. Ranganathan developed theoretical base for subject indexing based on theory of classification. He devised Chain Indexing system by using Colon Classification.

### **6.13 Types of Indexes:**

We need to know different types of Indexes. Broadly they are

- a. Author Index
- b. Subject Index
- c. Document Index
- d. Citation Index
- e. Cumulative Index
- f. Book Index
- g. Periodical Index
- h. Bibliographical Index
- i. Title index

### **6.14 Role of Indexing in Information Retrieval:**

1. Index acts as a link between a source of Information and its user.
2. If size of collection is quite large then Index plays a major role to retrieve relevant information. In Information Retrieval Systems, Index will guide or project itself as a guide to the concepts in a collection of documents.
3. A good Index minimizes the search efforts and ensures optimum results.
4. Index informs the existence of documents containing documents surrogated, such as author, title, imprint, and call number etc.

5. An Index is a systematic guide to concepts derived from a collection of documents represented by entries arranged in a known and searchable alphabetical, numerical classified order.
6. In Information Retrieval System, Index performs two simultaneous functions: -
  - a. Retrieving Information from documents that are required and
  - b. Holding back Information about required documents based on a particular subject.  
In the context of Information Retrieval System, the term Index is primarily used as a system capable of retrieving Information about required documents based on a particular subject.
7. The Two characteristics of Indexing i.e., exhaustivity and specificity affect two important measures of Information Retrieval System namely recall and precision, which operate the search stage or output stage of the system. Recall is the measure of the degree to which it delivers all relevant documents. Precision acts as filter in Information Retrieval System. It is measure of the system's ability to hold- back, unwanted item.
8. Indexing the concepts based on one of the indexing systems used as a tool, makes information retrieval possible.

### **6.15 Evolution of Indexing Systems:**

Towards the end of 15<sup>th</sup> century the practice of supplementing the systematic listing with an alphabetical subject index was introduced to increase the manipulative capacity of the catalogue. The catalogue which is an index to this store and the storing language had to face the strains in naming the new subjects in a natural way, problems of synonyms and homonyms and the inability in representing the documents with one term only. Though alphabetical subject index served as an index to a group, it did not specify or state more pointedly the subjects of the documents in a group. As a result uniformity in cataloguing entry was advocated. This led to the emergence of subject cataloguing and paved way for number systems put forth by various experts. Some of the major contributions have been discussed in the following pages.

**6.151 Cutter, C A:** First to establish a generalized set of rules for alphabetical subject headings in 1876 in his work "Rules for Dictionary Catalogue". He laid down several rules that went some way in solving the subject cataloguing problems. He preferred natural language as the only kind of terminology and emphasised on using the accepted names only. He was in favour of providing specific entry, putting the prominent word first and linking related subjects by cross references. These have been the basis of US practice in subject cataloguing even today.

**6.152 Kaiser, James:** Published "Systematic Indexing" in 1911. This work was an important step in the practice of alphabetical subject indexing. This was the first attempt to find a sound and consistent answer to the problem of significant order which is still valid and useful. Kaiser pointed out that many composite subjects can be analyzed into a combination of "Concrete and Process" of which "Concrete" is important than Process". For example: in "Chemical Treatment of Cancer" 'Cancer' is 'Concrete' and 'Chemical treatment' is 'Process'. Hence, the entry will be like: CANCER-Chemical Treatment.

Further, he introduced systematic arrangement of subheadings as opposed to straight alphabetical filing

**6.153 Coates, E J:** Most important contribution to the theory of alphabetical subject heading. In his book ‘Subject Catalogues’ he suggested that there should be a sound psychological basis for subject cataloguing. He categorized the compounds into twenty categories and these categories formed the basis for the significant order. For solving Subject- Space problem, he proposed ranking of subjects based on the nature of the subject. Some of the comments on this approach are: no natural use of languages, lengthy entries and complex but consistent and specific. These ideas were adopted in the British Technology Index.

**6.154 Ranganathan, S R:** Fundamental approach to deep rooted problem. Enunciated certain rules on the basis of which the subjects’ names could be framed. He developed a mechanical procedure known as ‘Chain procedure’ (you will learn much about this in Section 6.51). He utilized the concept of ‘Chain’ for deriving subject headings. In its notational representation a subject is translated back into a notational language in an order of hierarchy. In naming the subject the notation is translated back into terms digit by digit. The chain of terms from the first to the last digit stands for the full name of the subject. Necessity of Classification scheme, failures of links- missing, unsought, false, etc and of all the entries only one entry is specific; the others being generic are some of the flaws of the system.

**6.155 Metcalfe, J W:** Argued strongly for the alphabetic catalogue in his book “Tentative Code of Rules for Alphabetic Specific Entry published in 1957. He defined alphabetic – specific catalog as consisting of “Known names in known order”. He made important differentiation between ‘specification’ and ‘qualification’. Specification is division into species and Qualification is division by aspect, process or form. Argued that the purpose of subject cataloguing is to indicate the subject classes into which a document falls, not to indicate necessarily the precise subject of the term itself - Not to be Co-extensive, be Coordinate.

**6.156 Lynch, M F:** In his book “Articulated Subject Index” published in 1966, based his system on the preposition phrase. He intended to devise a method of generating subject indexes by computer manipulation of a simple sentence. These are of high standard. He treated prepositions as points.

### **6.16 Process of Indexing:**

The indexing process includes not only selecting terms to be indexed, but also adding qualifiers as subentries where appropriate, and editing the index after a first draft is produced to improve its cohesiveness, consistency, accuracy and usefulness to the reader.

The essential operations involved:

**1. Scanning the collection:** Indexer reviews the table of contents and introductory materials, skimming the entire publication to get sense of the topics and their

interrelationships, noting chapter headings and subheadings. (To get an over view of the document, to know what the document about)

**2. Analyzing the content:** The indexer tries to include in the index every significant term of Information in the document, choosing entries and subentries with an understanding of the way people will look up those terms in the Indexes. The indexer must understand the needs of the audience for the document, be thorough and consistent, know when to reword the author's thoughts (usually Index entries are created using words found in the text. But, sometimes indexer uses words not actually found in the text) and be able to tie related concepts together synonyms, etc) (See and See also references to connect related terms)

3. **Tagging** discrete items in the collection with appropriate identifiers, and

4. **Adding** to each identifier the precise location with in the collection where the term occurs, so that it may be retrieved.

The process of generating index entries calls for sensitivity to users' approaches, intuition to select appropriate terms and skill to identify relationship on the part of the indexer. The indexing process also requires the creation and recognition of pattern and rule consciousness and adherence to them as also accuracy and precision. In otherwords, the process involves the application of a model indexing system.

## **6.2 INDEXING SYSTEMS:**

An indexing system is a set of prescribed procedures for organising the contents of records of knowledge or documents for the purposes of retrieval and dissemination. An indexing system is the means whereby an indexing language can be applied to make an index. It may thus be stated that the need for an indexing system stems out of the work of devising index headings. These index headings mainly relate to documents dealing with compound or multi-concept subjects. For convenient study, indexing systems could be divided into two basic groups, the pre-coordinate systems and the post coordinate systems. A brief description of these is provided in the following paragraphs.

### **6.21 Pre-Coordinate Indexing system:**

Pre-coordinate indexing systems are conventional systems mostly found in printed indexes. In this type of systems a document is represented in the index by a heading or headings comprising a chain or string of terms. These terms taken together are expected to define the subject content of the document. The leading term determines the position of the entry in the catalogue or index, while the other terms are subordinated to it. For example, for the document entitled " The use of computers in Library and Information Activities" the following headings may be generated.

**Library activities:** Use of computers

**Computers:** Use in Library activities

Since the coordination of index terms in the index description is decided before any particular request is made, the index is known as a pre-coordinate index. Because this method of indexing coordinates the elements of compound subjects before any particular request is made for the information on that particular compound subject, it is known as pre-coordinate indexing. One of the characteristics associated with a pre-coordinate index

is that the headings in the index are relatively specific compared to one concept headings such as LIBRARIES or COMPUTERS. Pre-coordinate indexes are mostly prevalent as printed indexes. For example, the indexes relating to abstracting and indexing journals, national bibliographies apply principles of pre-coordinate indexing.

Two aspects are of great significance in relation to pre-coordinate indexes. The first aspect concerns the consistent description of subjects. In case of subject headings describing many concepts, consistency should be introduced into the terms used to represent individual concepts that constitute the multiple concept heading. Also, the order in which the individual terms representing the unit concepts of a multiple concept stated must be consistent. Some basic principles have to be evolved regarding an acceptable citation order of the terms. There must be a theoretical basis by which consistent citation orders could be achieved. Use of such theoretical principles may result in the derivation of a structural system of headings with consistent citation order between similar, yet distinct subjects.

The second significant aspect that requires the attention of subject cataloguers or indexers, is the need to provide access for those users who approach the subject under consideration for indexing from one of the secondary concepts. Since only one term can appear in the primary position in the prescribed citation order, the preferred citation order should be one, which caters to a majority of users. To help the remaining, references or added entries should be provided in the catalogue or index.

Alphabetical subject indexing systems devised by Cutter, Kaiser, Coats, Ranganathan (Chain Indexing), Metcalfe, Lynch, Sharp (SLIC- Selecting Listing in Combination), Fosket (Rotated Index), Craven (NEPHIS- Nested Phrase Indexing System), Austin (PRECIS- Preserved Context Indexing System), Bhattacharya (POPSI- Postulatebased Permuterm Subject Indexing), general classification schemes (Dewey Decimal Classification, Universal Decimal Classification, Library of Congress Classification, Colon Classification etc), library catalogues including OPAC's are the examples of pre-coordinate indexing systems.

### **6.22 Post-Coordinate Indexing Systems:**

These systems are also called coordinate indexing systems. The problems of pre-coordinate indexing specially the rigidity of citation order have led to the development of indexing techniques in a new direction, where the problem of taking a decision on the citation order is avoided by isolating the concepts of a composite subject and keeping them separate for manipulation at search stage. This device has shifted the coordination of index terms from input to output stage making the input work simpler and easier. As the coordination of terms takes place at the output stage, that is after the indexing operation itself, the system is called post-coordinate indexing. Post coordinate indexing systems can be grouped into two main categories: 'term records' and 'item records'. According to the former category, a concept is represented as a heading in a term card and accession numbers of all documents carrying that concept are posted on it. Thus we need to make as many entries for a document as we select terms to describe its contents. In each of these entries accession numbers of the documents will be available in

appropriate columns. Uniterm indexing of W E Batten is an example for this. According to the other approach, only one entry is made for a particular document and all aspects of the document are coded on the card. The mechanism followed here provides multiple access to the document through the coded concepts. The method is known as item entry system. Zatcoding of Calvin Moore is an example for this. Based on the basic principles of post-coordinate indexing a number of systems have been developed. The Uniterm system of Mortimer Taube, Optical coincidence method, Batten system, Selects system, Peek-a-boo system, Aperture card system, Peep hole card system, Edge-notched cards and Zatcoding are the examples of post-coordinate indexing systems.

**6.23 Titlebased Indexing System:** Post-coordinate systems developed as tools to overcome the inherent problems of pre-coordinate indexing systems, too suffer from few hidden problems. As a result much thought was given to overcome these problems including vocabulary control, establishing links, and weightage of the terms etc. There is of course one part of a document in which the authors usually try to define the subject: the title. A title is usually considered to be one line abstract of the document. In many cases this will give a clear indication of what the document is about unless the title of a document is fictitious. If we consider titles given to non-fictitious works, we will often find that works on the same subject have titles containing same significant words – keywords, which can be used as a basis for information retrieval. The use of keywords to give various kinds of index is well established, but has been emphasized in recent years by the use of computers to manipulate the terms. Catchword indexing, Keyword in context indexing, Keyword out of context indexing are the examples of this category.

### **6.3 INDEXING LANGUAGE:**

One of the important functions of information processing is to specify the subject of a document. This information is available in different parts of a document. The title of a book or an article or a technical report must succinctly indicate the subject content. But in some instances the titles are fictitious with catchy words. Abstract of the document in case of articles, and preface, introduction, contents page and book jacket in case of a book are also the sources of information regarding the subject of that document. Browsing through the whole document will help in ascertaining the subject of the document. Whatever may be the source of information, once the subject of a document is ascertained, it has to be recorded. At this stage, the role of language is vital. This language can be natural or artificial. Obviously this can be called as indexing language. The phrase “indexing language” is generally defined as all the words permitted either to describe a specific document or to construct a query to search a document file along with the rules describing how the terms are to be used and in what relation to each other. In other words, an indexing language is the language used to describe the subject or other aspects of information in an index or in a library catalogue. An indexing language is a "language" used for subject classification or -indexing of documents. (Not used about systems for descriptive cataloging or -indexing).

### **6.31 Types of indexing language**

Indexing languages have been categorized into a number of fundamental types. Indexing languages may be divided into (Fig 1) "classification systems" and "verbal indexing languages", although this is a superficial distinction. Classification systems may be divided into enumerative systems and faceted systems. Verbal indexing systems may be divided into controlled vocabularies and free text systems. Controlled vocabularies may be divided into pre-coordinate indexing systems and post-coordinate indexing systems. Indexing languages are kinds of metadata. Their function is to serve as subject access points (or to supplement other kinds of subject access points, e.g. references, cf., citation indexing)

From the theoretical perspective indexing languages may either be controlled systems or non-controlled systems. The non-controlled systems may further be divided into natural language or free text (Fig 2).

#### **6.311 Natural language indexing language**

In this approach any term from the document in question can be used to describe the document. But one of the greatest hurdles in discussing natural indexing languages is that it is not easy to identify or to know what exactly constitutes natural indexing languages. We do not generally come across lists of natural indexing languages. Obviously, a natural indexing language is the language of the documents that are indexed or catalogued for a library. Hence, it could be static as long as the document collection remains static. As soon as new batch of documents is added to the library, the terms of the indexing language change to accommodate the new terms contained in the new batch of documents. Another difficulty is that since each library or system indexes separate set of documents, each system will have a different indexing language even if the documents cover the same subject. Also, since the indexing language is derived from the documents, added to the library or input into the system, different records, even if they represent the same documents, may generate a different indexing language. These variations affect the consistency associated with the library catalogue and so, present many problems. Most of the natural indexing language is based upon the language of title, abstract and other text of documents.

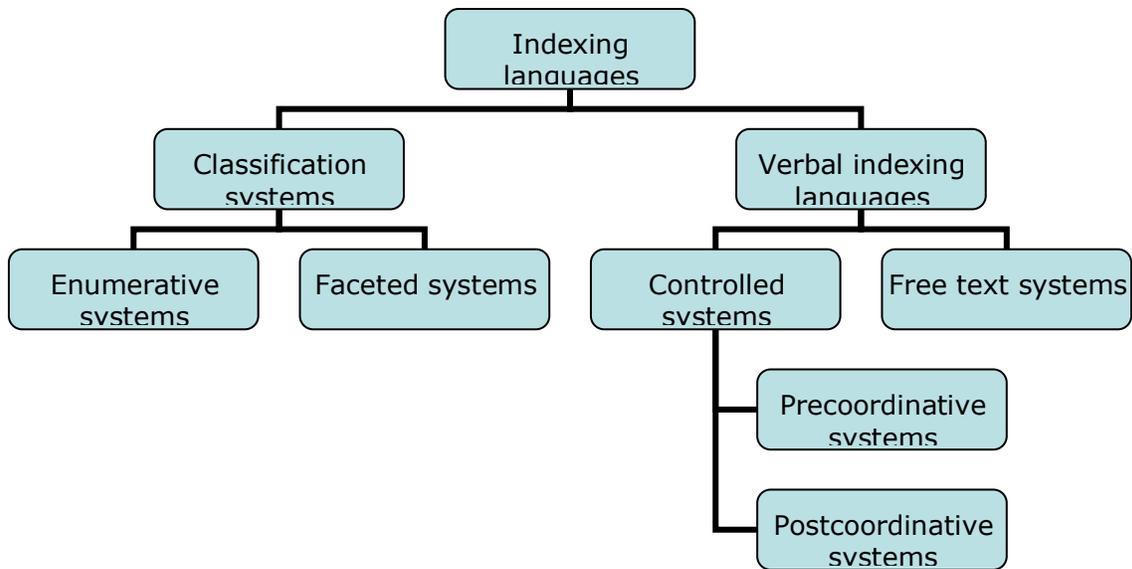


Fig. 1: Traditional view of the kinds of indexing languages

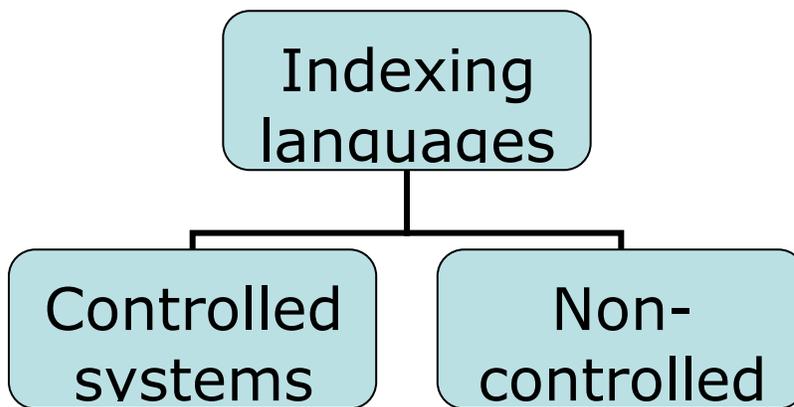
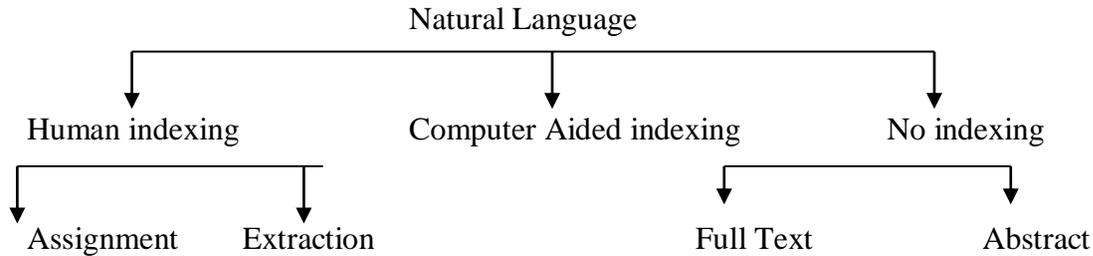


Fig. 2: Theoretically based view of the kinds of indexing languages

There is an active debate as to the effective use of natural language for indexing and subject cataloguing purposes. One school of thought believes that full exploitation of the opportunities offered by computer systems could only be done taking recourse to natural language indexing, where as the other school holds the view that controlled language indexing is the only proper way to index documents. Since the terms are chosen from the text itself, this approach may also be called indexing by extraction. One of the applications of natural language indexing is the production of indexes based on words in titles of documents such as KWIC and derived indexing systems. The Uniterm systems

derived in the early days of information retrieval are the examples of natural language systems.



### 6.312 Free indexing languages:

It is the nature of the free indexing language that any word or term (not only from the document) that suits the subject can be used to describe the document as subject heading or as an indexing term. The terms may be human assigned or computer assigned although free language indexing is commonly used in the computer produced catalogues or indexes. The computer operates by indexing under every word with which it is provided unless it is instructed not to do so. Hence free text indexes have high exhaustivity (one can search on every word). So it has potential for high recall, but will have much lower precision.

### 6.313 Controlled indexing language

Only approved terms can be used by the indexer to describe the document. These are often claimed to improve the accuracy of free text searching, such as to reduce irrelevant items in the retrieval list. These irrelevant items are often caused by the inherent ambiguity of natural language. Take the English word *football* for example. *Football* is the name given to a number of different team sports. Worldwide the most popular of these team sports is Association football, which also happens to be called soccer in several countries. The English language word football is also applied to Rugby football (Rugby union and rugby league), American football, Australian rules football, Gaelic football, and Canadian football. A search for *football* therefore will retrieve documents that are about several completely different sports. Controlled vocabulary solves this problem by tagging the documents in such a way that the ambiguities are eliminated. Compared to free text searching, the use of a controlled vocabulary can dramatically increase the performance of an information retrieval system, if performance is measured by precision (the percentage of documents in the retrieval list that are actually relevant to the search topic).

In some cases controlled indexing can enhance recall as well, because unlike natural language schemes, once the correct authorised term is searched, one need not worry about searching for other terms that might be synonyms of that term. However, a controlled indexing search may also lead to unsatisfactory recall, in that it will fail to retrieve some documents that are actually relevant to the search question. This is particularly problematic when the search question involves terms that are sufficiently tangential to the subject area such that the indexer might decide to tag it using a different term (but the searcher might consider the same). Essentially, this can be avoided only by

an experienced user of controlled indexing whose understanding of the the vocabulary coincides with the way it is used by the indexer.

Another possibility is that the article is just not tagged by the indexer because indexing exhaustivity is low. For example an article might mention football as a secondary focus, and the indexer might decide not to tag it with "football" because it is not important enough compared to the main focus. But it turns out that for the searcher that article is relevant and hence recall falls. A free text search would automatically pick up that article regardless. Controlled indexing systems are also quickly out-dated and in fast developing fields of knowledge, the authorised terms available might not be available if they are not updated regularly. Even in the base case scenario, controlled language is often not as specific as using the words of the text itself. Indexers trying to choose the appropriate index terms might mis-interpret the author, while a free text search is in no danger of doing so, because it uses the authors own words. The use of controlled vocabularies compared to free text searches are so costly because it requires human experts to index (or expensive automated systems), and the user has to be familiar with the controlled vocabulary scheme to make best use of the system. But as already mentioned, the control of synonyms, homographs can help increase precision.

## **6.4 CHARACTERISTICS OF INDEXING LANGUAGE:**

### **6.41 Purpose:**

Indexing language is designed for a special purpose. Other than subject description, indexing language is also used in organising a searchable file to be used by the information seekers. A match between the description of document done by the indexer and description of request made by the user will give positive result. This matching is possible if the file is organised in a predetermined order and the users are aware of it. In successful organisation of this file and subsequent matching of document and request, semantic and syntax of indexing language have very important roles to play.

### **6.42 Vocabulary Control:**

Vocabulary of indexing language must be precise and exact. A complete one to one relationship between the concepts and terms should be established. Synonyms, homonyms, homographs etc are controlled in indexing language. Out of the synonyms, only one term is accepted and this is used in subject description and file organisation. References are made from rejected terms to selected ones facilitating search by the users. In subject indexing this mechanism is provided with the help of "see" references. In a thesaurus this facility is provided with the help of "Use" reference. In a classification scheme synonyms are guided to the notation allotted in the scheme for classification.

### **6.43 Coordination of concepts:**

Specifying a concept with a particular term does not solve the problems in indexing a document having multiple terms. When multiple terms are used proper representation of subject depends on correct coordination of terms. In a natural language, apart from substantive words, a number of other auxiliaries like prepositions, conjunctions,

the meaning of the construction as a whole are made use of. But in an indexing language such auxiliaries are not available and as such the correct meaning of the subject or index heading has to be expressed largely through the order of words. A subject class generated by coordination of two or more terms representing different concepts, will differ from the classes represented by the individual terms or by the terms in some other combination. For example, the terms 'Library', 'School', and 'Building' may be coordinated in different ways like, 'Library school building', 'School library building', 'Building school library' and 'Building library school'. The meaning of each coordinated form differs from the other. In indexing language this coordination of terms is made to generate index phrase at input or output stage. It may be mentioned here that the coordination of terms is carried out at the input stage in pre-coordinate indexing and at output stage in post-coordinate indexing.

#### **6.44 Multiple Access:**

A subject representation is formulated with the help of the rules of syntax. To provide access from other terms, indexing languages provide some rules of rotation of component terms. The rotation is done in such a way that each of the component terms is placed in the access position, which is followed by other terms showing its context so that the correct meaning of the subject is represented. These rules of rotation are actually called special rules of syntax.

#### **6.45 Syndetic devices:**

Indexing language is an artificial one. Because of this artificiality the user needs guidance how to use it. This guidance to users is provided in indexing language with the help of various types of syndetic devices. These include guides, cross references, glosseries, inversion of headings, introduction to indexes etc.

**6.451 Guides and introduction to indexes:** Guides and introduction to specific indexes provide useful information regarding scope and structure of the index, terminology, and symbols used, rendering of subject headings, depth of indexing, format and typography, exceptions etc. Thus all important decisions taken by the indexer and editor are available here and a glance through this may save time and energy of the user.

**6.452 Cross references:** Cross references correlate similar concepts scattered throughout the index due to its alphabetical overtone. Also cross references guide the user from his present position in the index to where he should or should also be locate his desired information. These cross references are mainly of two types – 'see' and 'see also' or 'use' and 'used for'. Some examples are given below

Index Terms *see* Headings

College Library *see also* Academic Library

England *Use* Britain

Grain Crops *Used for* Barley

**6.453 Glossery:** The extension of the preferred term is sometimes limited or expanded to ensure a definitive coverage. Scope notes, explanations, enumeration etc indicate this. For example:

Cell Physiology: Structural property of cell

**6.454 Inverted Headings:** The normal order of words in a heading is sometimes inverted to bring the potent word in the beginning. Since users approach is expected by that word or to avoid scattering of related material. For example;

Bridges, Concrete  
Rubber, synthetic

**6.46 Relation Manifestation:**

Elementary terms of every language are related to each other paradigmatically (semantically) and syntagmatically (syntactically). This is true for indexing languages as well. Paradigmatic relations are those which are known in advance before scanning any particular document, while syntactic relations are understood only after scanning a particular document

**6.461 Paradigmatic relations:** Paradigmatic relations include broader, narrower relation and associative relation. These are often indicated by the arrangement of terms such as ;

EX: Chemistry  
    Organic Chemistry  
    Inorganic Chemistry  
        Metals

The organic chemistry and inorganic chemistry are parts of Chemistry; organic and inorganic chemistry are of same level and metals are part of inorganic chemistry. Users know these relations without any reference to a document. This relation is known in different ways in different languages.

**6.462 Syntagmatic relation:** Syntagmatic relation is restricted to a particular document. The relation is established between different concepts covered in the document only after the analysis of the subject content. For example, a document entitled “Law of inheritance and social status of women in rural India” has combination of concepts from ‘Law’, ‘Sociology’ and ‘Geography’, which is a special combination not normally shown in any indexing language. They present syntagmatic relation in the context of a specific document. To manifest this relation and indexing language has to devise some mechanism.

**6.47 Structural Presentation:**

Since an indexing language aims at efficient retrieval of documents. It has to consider the characteristics of users’ approaches. A user is normally interested in a particular subject but in the event of non-availability of any document on the specific subject or his

inability to decide about the exact extension of the subject of search or for collecting comprehensive literature on a topic, he may require materials on broader, narrower and collateral subjects. An indexing language has to therefore structure and display the relationship in a systematic manner. This is done in different ways by different types of languages. Thus indexing languages are structured languages.

#### **6.48 Effectiveness:**

The effectiveness of an indexing language can be measured only on the basis of the performance of the indexing system using that language. Expressiveness, ambiguity, compactness and the cost of usage are the important factors adding to the effectiveness of the system.

**6.481 Expressiveness:** This is the ability of the language to identify a subject, to distinguish between fine differences in subjects and to describe differing levels of details. This is best achieved by a natural language, but it is unsuitable for indexing work for various reasons. If we consider different language from this point of view, these are likely to fall in the following order of describing expressiveness: Faceted classification, Classaurus, Thesaurofacet, Thesaurus, Subject Headings List, and Enumerative Classification.

**6.482 Ambiguity:** By their nature, indexing languages are less ambiguous than a natural language, since ambiguity depends on synonyms and homonyms. If we consider different language from this point of view, these are likely to fall in the following order of decreasing expressiveness: Faceted classification, Classaurus, Thesaurofacet, Thesaurus, Subject Headings List, and Enumerative Classification.

**6.483 Compactness:** The compactness depends on the amount of information denoted by each term. On this count, the natural language seems to be least compact. In the syntactic language, as syntax approaches the complexity and variability of natural language, the amount of information per term begins to increase. Therefore, the different indexing languages may fall in the following order of increasing compactness: Enumerative Classification, Subject Headings List, Thesaurus, Thesaurofacet, Classaurus, and Faceted classification.

**6.484 Cost of Usage:** Cost here refers to the cost of training in the use of a language, the cost of indexing and retrieval and the cost of error rectification. The estimation of the cost is difficult. The different indexing languages may fall in the following order of decreasing cost of usage: Enumerative Classification, Subject Headings List, Thesaurus, Thesaurofacet, Classaurus, and Faceted classification.

### **6.5 INDEXING MODELS:**

You have learnt in this Unit from Section 6.1 to 6.484, the concept of index, indexing, indexing system, indexing language, and characteristics of indexing languages. In this section you will be introduced some of the major subject indexing models, namely: Chain Indexing, PRECIS and POPSI and KWIC – a titlebased indexing model.

### **6.51 Chain Indexing:**

Ranganathan designed a new method of deriving verbal subject headings to provide subject approach to documents through the alphabetical part of a classified catalogue. This method was different from the enumerated subject headings systems like Library of Congress Subject Headings and Sears List of Subject Headings. This is a mechanical method to derive subject index entries or subject headings from the Class Number of a document, called chain procedure. Ranganathan defined it as a “*Procedure for deriving Class Index Entry (i.e Subject Index Entry) which refers from the name of a class to its class numbers in a more or less mechanical way. A note is also Class Index Entries in a classified Catalogue and Specific Subject Entries, Subject Analyticals, and see also Subject Entries in a Dictionary Catalogue*”. This method may be used to provide indexes not only to classified catalogues and classification schemes, but also to other systematically organized indexes, even when they are arranged alphabetically.

#### **6.511 Steps in Chain Procedure:**

1. Determination of the specific subject by analyzing the subject content of the document
2. Representation of the name of the specific subject in terms of its fundamental components removing all auxiliary words from the title
3. Determination of the category or status or role of each fundamental categories according to a set of principle and postulates
4. Transformation of the analyzed name of subject by rearranging if necessary, fundamental components according to a few additional postulates and principles
5. Standardization of each term in the transformed name of the subject
6. Translation of the name of the subject of the document in standard terms into its class number in the preferred classificatory language
7. Determination and representation of the chain of which the name of the specific subject is the Last Sought Link
  - 7.1. Representation of the class number in the form of a chain in which each link consists of two parts – the class number and its translation – resulting in the name of the subject in standard terms, which may be done according to the following procedure:
    - Make the first link from first digit
    - Make the second link out of two digits and so on up to the last link which is to be made of all digits
    - Write the links one below the other in succession
    - Write against each link its translation into natural language
    - Connect each link with its translation by an “=” sign and
    - Join “=” sign of each link with that of the next succeeding link by a downward arrow if necessary
8. Determination of the different kinds of links, Ex: False Link, Unsought Link, Sought Link and Missing Link
  - 8.1. False Link: A link which is
    - Not a class number

- The last link of a compound class number and does not have a name in a natural language Ex: A link is a False Link if it ends with a connecting symbol or digit representing phase relation or time isolate idea representing time itself
- 8.2. Unsought Link: A link which
    - Ends with a part of the isolate focus in a class number
    - Represents a subject on which reading material is not likely to be produced or sought or which is not likely to be looked up by any reader seeking materials on specific subject forming last link of the full class number
  - 8.3. Sought Link: A link which is neither false nor unsought
  - 8.4. Missing Link: A link in a chain with gap corresponding to the missing isolate in the chain
  9. Deriving of a subject heading from each of the sought links in the chain, according to a set rules formulated for the purpose in view
  10. Construction of subject heading for specific subject entry or subject reference entry
  11. Recording the subject
  12. Providing Cross - references

The rules and procedures of chain indexing can be explained with the following example.

**O,111,2J64, 11 : Othello**

- O : Literature (Sought link)
- O, : (False link)
- O, 1 : Indo-European Literature (Unsought Link)
- O, 11 : Teutonic Literature (Unsought link)
- O,111 : English literature (Sought link)
- O,111, : (False link)
- O,111,2 : English drama (Sought link)
- O,111,2J : English drama during 15<sup>th</sup> century (Unsought link)
- O,111,2J6 : English drama during 1560's (Unsought link)
- O,111,2J64 : Shakspere (Sought link)
- O,111,2J64, : (False link)
- O,111,2J64, 11 : Othello (Sought link) [Note: Othello is assumed to be the first work]

The above chain generates the under mentioned subject headings and Class Index Entries

- Othello, Shakespere, Drama, English literature, Literature : O,111,2J64,11
- Shakespere, Drama, English literature, Literature : O,111,2J64
- Drama, English literature, Literature : O,111,2
- English literature, Literature : O,111
- Literature : O,111

**6.512 Comments:** Chain indexing is a systematic, consistent and almost mechanical method of deriving subject entries. With its postulational approach and principles, it has a strong theoretical foundation of classification. The procedure economizes on the number of subject entries and provides for exhaustive pinpointed retrieval efficiency. It is a useful method not only for deriving subject entries but also for retrieval in bibliographies and micro level documents. Its classificatory method makes it possible for deriving subject entries for documents in languages other than English. Subject headings can be derived from the verbal chain. The procedure's reliance on classificatory principles is its strength as well as weakness. It operates quite well with a faceted classification scheme like Colon Classification. With enumerative systems (like DDC) the efficiency suffers. With all these problems the Chain indexing is a powerful method of subject indexing with its potential for further refinement and efficiency.

### **6.52 Preserved Context Indexing System (PRECIS):**

PRECIS stands for Preserved Context Indexing System. It was designed and developed by Dereck Austin by about 1970. It was intended to provide a new system of subject indexing for the British National Bibliography which was launching the UK/MARC Project. Any entry in this subject index system retains the full context of the term of approach or the term sought by the users of an information system. PRECIS has been defined by its author Austin "as a system in which the initial string of terms, organized according to a scheme of role indicating operators is computer manipulated so that selected words function in turn as the approach term".

Entries are restructured at every step in such a way that the user can determine from the layout of entry, which terms set the appropriate terms into its context and which terms are context dependent upon the approach term. PRECIS has two levels of operation. In the first level (human level), subject statement is analyzed into a set of roles and in the second level (computer level), the analyzed subject statement is programmed to be manipulated into producing a variety of PRECIS subject index entries by computer processing and printout.

**6.521 Theoretical Framework:** The basic problem in any indexing system is to control the terminology scatter (synonym/homonym) and the other is synthetic scatter (same message by different set of expression). PRECIS is able to control these by reference system and syntactic control system. The PRECIS index entry format is a two line three parts style as shown below.

**Lead                      Qualifier**

**Display**

Lead and Qualifiers are given the first line and other contextual terms providing access to the statement in the second line. PRECIS is also a version of Chain indexing. Here the Chain is referred to as a string. It is a string of role operators. The syntactic control is essentially in the form of role operators. It is governed by two principles for organization – one 'context dependence' and the other 'one to one relationship'. This enables PRECIS to analyze the subject matter of a document. The role operators regulate the writing of

conceptual terms and also regulate the sequence of terms. In order to set down the selected indexing terms from a document in the sequence of context dependency and to ensure that a term is consistently placed at the input order a “Schema of role operators”(Fig 3) is used. This schema determines the order in which terms should be cited and hence could be regarded as a kind of indexing grammar. Through the use of this schema, indexer expresses the relationships between the terms in a summary statement.

**Fig 3: Schema of Role Operators**

<b>Primary Operators</b>		
Environment of Core Concepts	0	Location
Core Concepts	1	Key System Thing when action not present. Thing towards which an action is directed
	2	Action: Effect of action
	3	Performer of transitive action (agent, instrument): Intake factor
Extra-core concepts	4	View point as form
	5	Selected instance: Study region: Study example: Sample population
	6	Form of Document: Target user
<b>Secondary Operators</b>		
Coordinate concepts	f	Bound coordinate concept
	g	Standard coordinate concept
Dependent elements	p	Part: Property
	q	Member of quasi-generic group
	r	Assembly
Special classes of action	s	Role definer: Directional property
	t	Author attributed action
	u	Two way interaction

<b>Primary Codes</b>		
Theme Inter-links	\$x	1 <sup>st</sup> concept in coordinate theme
	\$y	2 <sup>nd</sup> / subsequent concept in coordinate theme
	\$z	Common concept
Term codes	\$a	Common noun
	\$c	Proper name
	\$d	Place name
<b>Secondary Codes</b>		
Differences		
Preceding differences (3 Characters)	1 <sup>st</sup> and 2 <sup>nd</sup> characters	
	\$0	Non-lead space generating
	\$1	Non-lead close up
	\$2	Lead space generating
	\$3	Lead close up
	3 <sup>rd</sup> character = number in the range 1 to 9 indicating level of difference	
Date as difference	\$d	
Parenthetical difference	\$n	Non-lead parenthetical difference
	\$o	Lead parenthetical difference
Connectives	\$v	Downward reading connective
	\$w	Upward reading connective

Typographic codes	
\$e	Non-filing part in italic preceded by comma
\$f	Filing part in italic preceded by comma
\$g	Filing part in Roman, no preceding punctuation
\$h	Filing part Italic preceded by full point
\$I	Filing part in italic no preceding punctuation

### 6.522 Format

The entries generated can be displayed in three different formats, Standard, Predicate transformation and Inverted.

**6.5221 Standard Format:** In this format lead and qualifiers appear in the first line and the contextual terms in the second line. The term in these sections is derived by the process of shunting of terms from the string as indicated in the document. This can be called Lead-Qualifier-Display format or standard format.

If there are four terms A, B, C, D, the positions these occupy in the Lead, Qualifier and Display in the different entries would be as follows.

A  
BCD

B     A  
CD

C     BA  
D

D     CBA

EX: The standard entries for the input string given below are as follows

(0) India. (1) Textile industries, (p) Personnel. (2) Recruitment

INDIA

Textile industries, Personnel.Recruitment

TEXTILE INDUSTRIES.     India

Personnel.Recruitment

PESONNEL. Textile industris.India

Recruitment

RECRUITMENT.     Personnel. Textile industries. India

**6.5222 Inverted Format:** Whenever a term coded (4), (5) or (6) occurs in a string of conceptual terms in the leading position, this appears in bold and the dependent element in italics.

EX: The indexentries for the input string given below are as follows

(0) India. (1) Schools (6)Directories

INDIA.

Schools. *Directories*

SCHOOLS. India

*Directories*

DIRECTORIES

India. Schools

**6.5223 Predicate Transformation:** The format is produced whenever a string containing a term that represents an agent code (3) comes as lead or term prefixed by one of the operators (2, s, t) which are indicative of some kind of action. This prevents the anomaly of a term followed by another term in the display in one case and in the qualities in another. As the action and key system together form the predicate it is named as Predicate transformation

EX: The index entries for the input string given below are as follows

(0) India. (1) Crops (2) Damage \$v by \$w to (3) Droughts

INDIA

Crops. Damage by droughts

CROPS. India

Damage by droughts

DAMAGE. Crops. India

By droughts

DROUGHTS. India

Damage to Crops

### **6.523 Syntax, Semantics, Vocabulary and Cross references:**

**6.5231 Syntax:** The syntax of the system is governed by the principle of context dependency. This is achieved by tagging role operators to individual terms in the string. This is a posteriori or document dependent relationship. Relationships of concepts are reflected in the input string. In a linear input string this relation is reflected in the following manner.

India → Tobacco industries → Personnel → Training

**6.5232 Semantics:** Semantic relationships are a priori or independent of documents. The semantic aspects of this system are conventional. In addition to the 'Genus-Species' relation, equivalence and associative relationships between concepts are projected with the help of 'See' and 'See also' references.

**6.5233 Vocabulary and Cross references:** The vocabulary used in PRECIS is controlled, structured and open ended. As a controlled vocabulary out of the synonyms a single preferred term is used to represent a concept. As a linking device 'see' references are provided between the preferred term and synonyms. The structuring of vocabulary takes care of different semantic relationships that exist between different terms. The currency in vocabulary is achieved by open-endedness.

**6.524 Comments:** PRECIS is one of the most successful indexing systems. One of the reasons for its outstanding success is the support it got from British National Bibliography. It now forms part of MARC records generated in many national bibliographies. Its theoretical foundation, initiated and supported by the Classification Research Group of Great Britain is another highlight for its success. The PRECIS manual contributed very considerably to its application in many situations. PRECIS is criticized as expensive system to use and is complex to learn and use. It is inefficient if used manually, as the generation of entries is a time consuming job. The application of the role operators for the formation of the string may not produce the same results when different indexers interpret the operators with reference to the different concepts. With all these limitations PRECIS claims to be one of the best systems of indexing; best among the contemporary indexing models.

### **6.53 Postulate-based Permuted Subject Index (POPSI):**

Postulate-based Permuted Subject Index (POPSI), designed and developed by Ganesh Bhattacharya of the Documentation Research Training Center, Bangalore, is another indigenous indexing model besides Ranganathan's Chain indexing. POPSI can be applied to micro and macro level documents available in the form of non-print/non-books forms. Steps involved in the process make it clear that the system has rightly been named. The process is completed in several steps as follows: analysis, formalization, standardization, modulation, preparation of entries for organizing classification, decision about terms of approach, preparation of entries for associative classification and alphabetization. The first six steps are based on certain principles and the seventh is based on the techniques of permutation.

### **6.531 Theoretical Foundations:**

POPSI is not based on any particular system of classification but built around a set of fundamental theoretical ideas on classification both in the analysis of subjects as well as in the structuring of the names of subjects. The deep structure of POPSI arises from a Subject Indexing Language, which should form the basic framework for any system of subject indexing. All ideas concrete or conceptual could be regarded as a

manifestation of one or the other of a set of postulated ‘Elementary Categories’ of POPSI. These elementary categories are: Discipline (D); Entity (E); Action (A); Property (P) and Modifier (M)

**D= Discipline:** An elementary category that includes conventional fields of study or any aggregate of such fields. Ex: Physical Science, Physics, Chemistry etc.

**E= Entity:** An elementary category that includes manifestation having perceptual correlation or only conceptual existence as contrasted with their properties and actions performed by them or on them. Ex: Energy, Light, Plants, Time, Environment etc.

**A = Action:** An elementary category that includes manifestation denoting the concept of doing action which may manifest as self-action or external action. Ex: Function, Migration, Selection, etc.

**P=Property:** An elementary category that includes manifestations denoting the concept of attribute – qualitative or quantitative. Ex: Property, Effect, Power, Capacity, Efficiency, Utility etc.

**M= Modifier:** In relation to a manifestation of any one of the elementary categories D, E, A and P the term Modifier refers to an idea used or intended to be used to qualify the manifestation without disturbing the conceptual wholeness of the latter. With the help of a modifier extension of a qualified manifestation is decreased and the intention is increased. Thus ‘Sanskrit’ in ‘Sanskrit drama’ is a modifier.

### 6.532 Syntax, Semantics and Sequencing of terms:

**6.5321 Syntax:** The syntax of the system is based on the principles and postulates of general theory of classification propounded by Ranganathan. Precise citation order of the components of subject formulation is achieved with the help of numerical devices available with POPSI table (Fig 4). In earlier versions of POPSI punctuation marks and inclusion symbol (>) were used to perform this function. But the latest version of POPSI has not discarded the punctuation marks completely and punctuation marks like comma (,), full stop (.) and hyphen (-) have their distinct role.

**Fig 4: POPSI Table**

0 Form Modifier	6 Entity
1 General treatment	

2 Phase relation	7 Discipline
2.1 General	.1 Action , Part
2.2 Bias	.2 Property . Species/Type
2.3 Comparison	- Special Modifier
2.4 Similarity	
2.5 Difference	Note: As .1 and .2 are dependent ECs, these are to be prepared by the notations for the ECs to which these are dependent.
2.6 Application	
2.7 Influence	
Common Modifiers	8 Core
3 Time	9 Base [Features analogues to 6/7]
4 Environment	
5 Place	

**6.5322 Semantics:** The system requires a mechanism for vocabulary control characterized by special features. A hybrid of faceted classification and thesaurus has been suggested for this purpose known as “classarus”. Within the systematic part terms are displayed under seven categories accompanied by all the synonyms and subordinates of successive orders. Alphabetical index part incorporates all the terms from systematic part along with their addresses. POPSI also permits the use of prepositions, conjunctions, participles, etc whenever necessary to convey exact meaning and to avoid ambiguity.

**6.5323 Sequence of terms:** The sequence of component terms in the subject propositions of this system is governed by the syntax of POPSI. According to general principles of sequencing D is followed by E (modified or otherwise) interpolated or extrapolated by A and/ or P (either modified or otherwise). To indicate the categories of component terms and to fix their positions in subject formulations in precise manner the help of POPSI table is taken. A manifestation of action (A) or property (P) follows the manifestation in relation to which it is A or P. Similarly a part or a modifier (M) follows immediately the manifestation to which it is related. Punctuation marks hyphen (-), full stop (.) and comma (,) are used in that sequential order.

### **6.533 Steps in POPSI:**

Let us generate the index entries for the title “Chemical treatment of tuberculosis of lungs” with the help of POPSI. The steps involved in the POPSI procedure are as follows.

**6.5331 Analysis:** Subject indicative statement of the explicit proposition of the subject is analyzed to identify the facets in terms of concepts and modifiers. Analysis for the above example will lead to the following.

Medicine (D)  
Chemical treatment (A)  
Tuberculosis (P)  
Lungs (E)

**6.5332 Formalization:** In the stage of formalization the sequence of components derived by analysis has to be decided. The accepted sequence of the system is D, E, A, modified or unmodified, appropriately interpolated or extrapolated by P, modified or unmodified. Applying this principle the components are sequenced in the following manner.

Medicine (D), Lungs (E), Tuberculosis (P of E), Chemical treatment (A on P of E)

**6.5333 Standardization:** The third stage is concerned with the semantics of POPSI language. This step of standardization helps to decide the standard terms for synonyms and the terms for reference generation. In the present example 'Chemical treatment' may be substituted by 'Chemotherapy'. A vocabulary control mechanism is used to achieve this goal. This operation results in the standardized basic chain like the following:

Medicine (D), Lungs (E), Tuberculosis (P of E), Chemotherapy (=Chemical treatment) (A on P of E)

**6.5334 Modulation:** The fourth stage of operation augments the standardized subject formulation by interpolating or extrapolating the successive superordinates by using standard terms along with their synonyms. This step also involves semantic aspect of POPSI as vocabulary control is necessary here. On applying this operation we get the basic modulated chain:

Medicine (D), Man, Respiratory system, Lungs (E), Disease, Tuberculosis (P of E), Treatment. Chemotherapy (=Chemical treatment) (A on P of E)

**6.5335 Preparation of EOC:** This step leads to the preparation of the main entry, which is further used for generation of associative entries. In order to achieve this goal a systematic set of numbers as given in the POPSI table is used to indicate the categories and positions of the components in the subject formulation. The structure of the entry after this operation will look like this:

Medicine 6 Man, Respiratory system, Lungs 6.2 Disease. Tuberculosis 6.2.1  
Treatment. Chemotherapy

(Notation for manifestation of Discipline (D) is omitted in the basic version of POPSI)

**6.5336 Decision about terms of approach:** This step is concerned with the decision regarding terms of approach for generating successive index entries and references. Terms of approach are judiciously selected to see that economy is achieved without disturbing basic requirements. In this process of selection generalized terms like man, animal, etc are omitted from entry points. During this operation synonyms are controlled and references are generated from synonyms to standard terms. For the present example

all the terms other than 'Medicine' and 'Man' are selected as terms of approach and the following reference is generated.

Chemical treatment  
See Chemotherapy

**6.5337 Preparation of EAC:** This step arranges entries under each approach term. The complete subject formulation is repeated under all the approach terms from the index phrase

Respiratory system

Medicine 6 Man, Respiratory system, Lungs 6.2 Disease. Tuberculosis 6.2.1  
Treatment. Chemotherapy

Lungs

Medicine 6 Man, Respiratory system, Lungs 6.2 Disease. Tuberculosis 6.2.1  
Treatment. Chemotherapy

Disease

Medicine 6 Man, Respiratory system, Lungs 6.2 Disease. Tuberculosis 6.2.1  
Treatment. Chemotherapy

Tuberculosis

Medicine 6 Man, Respiratory system, Lungs 6.2 Disease. Tuberculosis 6.2.1  
Treatment. Chemotherapy

Treatment

Medicine 6 Man, Respiratory system, Lungs 6.2 Disease. Tuberculosis 6.2.1  
Treatment. Chemotherapy

Chemotherapy

Medicine 6 Man, Respiratory system, Lungs 6.2 Disease. Tuberculosis 6.2.1  
Treatment. Chemotherapy

**6.5338 Alphabetization:** Here all the index entries including references are arranged in a word by word sequence. The final display of index entries will look like this.

Chemical treatment

See Chemotherapy

Chemotherapy

Medicine 6 Man, Respiratory system, Lungs 6.2 Disease. Tuberculosis 6.2.1  
Treatment. Chemotherapy

Disease

Medicine 6 Man, Respiratory system, Lungs 6.2 Disease. Tuberculosis 6.2.1  
Treatment. Chemotherapy

Lungs

Medicine 6 Man, Respiratory system, Lungs 6.2 Disease. Tuberculosis 6.2.1  
Treatment. Chemotherapy

Respiratory system

Medicine 6 Man, Respiratory system, Lungs 6.2 Disease. Tuberculosis 6.2.1  
Treatment. Chemotherapy

Treatment

Medicine 6 Man, Respiratory system, Lungs 6.2 Disease. Tuberculosis 6.2.1  
Treatment. Chemotherapy  
Tuberculosis  
Medicine 6 Man, Respiratory system, Lungs 6.2 Disease. Tuberculosis 6.2.1  
Treatment. Chemotherapy

#### **6.534 POPSI-Specific:**

Basic version of POPSI can be manipulated to generate POPSI-specific to meet specific requirement of an information system. If we want to bring all or major part of information on “Tuberculosis” together irrespective of its occurrence in man or animal , POPSI- Specific is to be developed. In such a case “Tuberculosis” which is considered as a manifestation of property (P) according to the basic version, is considered as “Base” (B) in this case. If in this case property (P) is involved, it should be treated as the “Core” going with the “Base”. Once this decision is taken different steps of POPSI procedure are followed to derive the index entries.

#### **6.535 Comments:**

POPSI is based on extensive basic thinking to design a system that would combine the best features of classificatory principles and alphabetical subject indexing. The total history of the system has roots in Chain indexing and later in General Theory of Subject Indexing Languages”. POPSI is flexible in the sense that its basic version can be manipulated to produce POPSI-Specific to meet specific requirements. Though POPSI is basically designed for manual methods, it is also amenable to computers. The only and the genuine drawback with POPSI is, it does not enjoy international support unlike PRECIS.

#### **6.54 Keyword in Context Indexing System:**

The system is based on the usage of natural language terminology for deriving the index entries. Thus instead of concepts significant words, as they appear in the titles of documents, are used as keywords or catchwords. A simple KWIC index is based on the ‘keywords’ that appear in the title of the documents.

**6.541 Structure:** The index consists of three parts – keyword, context and identification code. The keyword which is displayed at the window provides access point to the search. All keyword are displayed in turn at the window and they appear in the index at the appropriate place according to alphabetical order. Index entries are generated in association with all of the words in the batch of titles. Excluding the keywords at the window other terms in the index line express the context in which the keywords have been used. Identification code at the extreme right indicates the location of the document.

Ex: RETRIEVAL/Automatic Information Organization - 17

Here, Retrieval is the keyword, Automatic information organization is the context and 17 is its identification number

**6.542 Format:** KWIC indexes comprise three parts – keyword, context and identification number. A sign (/ , = or anything else) is used to indicate the end of the title. For example a document entitled “Use of roles and links in indexing’ and with the identification number 16 will have the following entries.

<b>INDEXING/</b> Use of roles and links in	16
<b>LINKS</b> in indexing/ Use of roles and	16
<b>Roles</b> and links in indexing/ Use of	16

The format has some visual glaring as it has too white space. As a solution, the window or access line of the index is shifted from the extreme left to the center of the page and the title is wrapped around or recirculated along the index line as mentioned below.

and links in	<b>INDEXING/</b> Use of roles	16
of roles and	<b>LINKS</b> in indexing/ Use	16
Indexing/ Use of	<b>ROLES</b> and links in	16

**6.543 Variations of KWIC:**

Modification in the format of KWIC index could not offer proper satisfaction since the filing order was not at the normal place. Thus various printed formats have been experimented. KWOC (Keyword out of Context), KWWC (Keyword with Context), KWAC (Keyword and Context), KLIC (Key Letter in Context) and SWIFT (Selected Words in Full Titles) etc have been designed.

**6.544 Comments:**

All these title based indexes can be produced speedily, and no intellectual effort is required. These provide a useful current awareness service. These have their limitations. In particular, these fail unless titles are explicit. Moreover, lack of subject analysis, points to the fact that they cannot be an effective tool to retrieve the information. The alphabetical approach scatters information on a specific topic. For a large collection KWIC becomes a costly affair. In spite of the problems, these can be of definite help to users who can identify a document by its title.

**6.6. Check your progress**

1. PRECIS was developed by Derek Austin for use in the  
 (a) **BNB** (b) INB (c) ISBD (d) ISBN
  
2. POPSI is a  
 (i) post co-ordinate indexing system (ii) keyword based indexing system  
 (iii) **pre co-ordinate indexing system** (iv) citation indexing.

3. The concept of 'Stopword' list is relevant in the context of

(a) Uniform Indexing (b) Citation Indexing (c) Chain Indexing (d) **Keyword Indexing**

4. Uniterm is an example of

(a) **Post Co-ordinate Indexing** (b) Automated Indexing (c)

Subject Indexing (d) Pre Co-ordinate Indexing

5. Chain indexing is an example of

(a) Post Co-ordinate Indexing (b) Automated Indexing (c) Subject

Indexing (d) **Pre Co-ordinate Indexing**

### **6.7 Summary:**

In this unit you have been explained the concept of index and indexing. The functions/uses of index have been discussed. Further the history of indexing has been traced. Also, different types of indexes have been listed. The role of Indexing in Information Retrieval has been discussed. The evolution of Indexing Systems from Cutter to Lynch is traced. The process of indexing has also explained to you. You have also been explained the differentiation between Pre-Coordinate, and Post-Coordinate Indexing Systems. In this unit you also have studied the concept of Indexing Language, types of indexing languages and their characteristics. At the end you have studied the salient features of indexing models: Chain Indexing, Preserved Context Indexing System (PRECIS), Postulate-based Permuted Subject Index (POPSI) and Keyword in Context (KWIC).

### **6.8. Glossary**

- KWOC (Keyword out of Context),
- KWWC (Keyword with Context),
- KWAC (Keyword and Context),
- KLIC (Key Letter in Context) and
- SWIFT (Selected Words in Full Titles) etc are the variations of KWIC.

### **6.9 Questions for self study**

1. Describe the features of KWIC.
2. Mention the variations of KWIC
3. List the uses of indexes
4. Highlight the role of indexing in information retrieval
5. Briefly indicate the steps involved in the process of indexing

6. Write the features of pre-coordinate indexing systems
7. Write the features of post-coordinate indexing systems
8. Define an indexing language. Mention the different types of indexing languages
9. Write the features free indexing languages
10. What are syndetic devices?
11. What do you mean by relation manifestation?
12. How the effectiveness of an indexing language is measured?
13. Mention the steps involved in chain indexing
14. Explain the merits and demerits of chain indexing.

#### **6.10 REFERENCES:**

1. Chakraborty, A R and Chkraborthy, Bhubaneswar. Indexing: Principles, Process and Product. Calcutta: World Press, 1984
2. Fosket A C. Subject approach to information, Ed 5. London: Library Association, 1996
3. Ghosh, S K and Satpathi, J N. Subject Indexing Systems: Concepts, Methods and Techniques. Calcutta: IASLIC, 1998
4. Guha, B. Documentation and Information: Services, Techniques and Systems. World Press, 1983
5. Hutchins, W. J. Languages of indexing and classification. A linguistic study of structures and functions. London: Peter Peregrinus, 1975
6. Lancaster, F. W. Vocabulary Control for Information Retrieval. 2nd ed. Info Resources Press, 1986.
7. Lancaster, F. W. Indexing and abstracting in theory and practice. London: Facet Publishing, 2003
8. <http://www.marisol.com/>
9. Prasher, R.G. Index and Indexing Systems. Ludhiana: Medallion Press, 1989
10. Riaz, Mohammad. Advanced Indexing and Abstracting Practices. New Delhi: 1989

---

## UNIT 7: SUBJECT HEADINGS LISTS, THESAURUS, THESAUROFACET

---

### Structure

- 7.0 Objectives
- 7.1 subject Headings List:
  - 7.11 Library of Congress Subject Headings (LCSH)
  - 7.12 Sears List of Subject Headings
- 7.2 Thesaurus:
  - 7.21 Definition:
  - 7.22 What is in a Thesaurus?
  - 7.23 Construction of Thesaurus:
  - 7.24 Study of Thesaurus
    - 7.241 Medical Subject Headings (Mesh®)
    - 7.242 INSPEC Thesaurus:
- 7.3 Thesaurofacet:
  - 7.31 BSI Root Thesaurus:
- 7.4 Classarus:
- 7.5 Check your progress
- 7.6. Summary
- 7.7. Glossary
- 7.8 Questions for self study
- 7.9 References

## **7.0 OBJECTIVES:**

You have studied the various controlled vocabulary tools, their purpose and functions in Unit 5. In this Unit you will study

- ❖ The different subject headings lists like LCSH, Sears List etc
- ❖ The concept of thesaurus, its construction and two important thesauri, mesh and INSPEC
- ❖ The concept of thesaurifacet
- ❖ Study of BSI ROOT Thesaurus
- ❖ The concept of classarius

## **7.1 SUBJECT HEADINGS LIST:**

Subject access to information has traditionally been provided in one of two ways: through classification, and through assigning terms or phrases from a standardized vocabulary such as a list of subject headings or a thesaurus. Classification involves the development and use of a scheme for the systematic organization of knowledge. The second category is subject headings' Lists. These are used to assign subject headings describing the content of the object, being catalogued. Subject heading systems have various tools to assist cataloguers - for example paper or written manuals of rules or guidelines, scope notes which clarify how terms should be used and may draw attention to distinctions between terms, references. A number of subject headings lists have been designed and developed over a period of time. Many of these are in use in libraries and other bibliographical publications all over the world. By far the most widely used is the LCSH - Library of Congress Subject Headings and Sears List of Subject Headings. In this unit you shall study these systems.

### **7.11 LIBRARY OF CONGRESS SUBJECT HEADINGS (LCSH)**

Library of Congress Subject Headings (LCSH) list is the most comprehensive list and accepted as the worldwide standard. It provides an alphabetical list of all subject headings, cross-references and subdivisions in verified status in the LC subject authority file. It is the official subject-heading list of the Library of Congress, the national library of USA. The LCSH comprises a thesaurus of subject headings, maintained by the United States Library of Congress (LC), for use in bibliographic records. The first edition was published in two volumes between 1909 and 1914 as the "List of Subject Headings used in the Dictionary Catalogues of the Library of Congress", though the work of compilation started way back in 1897. Supplements were issued at regular intervals until the publication of the second edition in 1919. Since then, the list has grown steadily and constantly updated and revised. Presently it is in its 29<sup>th</sup> edition (2006). This edition contains over 280,000 total headings and references. The basis for the LCSH was devised largely by J C M Hanson, Chief of the Catalog Department from 1897 to 1910. With the computerization of the library's bibliographic operations the LCSH became computer-based publication in 1986 with computer produced 10<sup>th</sup> edition of LCSH.

**7.111 Functions:** The Library of Congress subject headings system serves as a controlled vocabulary for subject indexing of the Library of Congress collection and for use in subject cataloging and indexing by other libraries or indexing agencies

**7.112 Fundamental principles:** The fundamental principles guiding the development of the Library of Congress subject headings system are effective responses to user needs and to literary warrant, use of uniform headings (one heading per subject - control of synonyms), unique headings (one subject per heading - control of homographs), and provision of direct access to specific subjects, stability, and consistency.

**7.113 Formation of Subject Headings:** Subject headings are constructed in LCSH in a variety of ways, ranging from lists of single words representing unit concepts and descriptive phrases of single concepts in two or more words to complex and compound subject dealing with combinations of aspects or facets of subjects and different types of phase relations between subjects. Let us study these headings

**7.1131 Single word headings:** There are documents that deal with a single subject or topic which can be represented by a single word. Usually a single word represents objects, things, persons or concepts that can be precisely expressed by such single words.

For example:

Economics                  Forgery                  Humanism                  Railways                  Advertising

Generally there should be no ambiguity in the meaning of such single concepts that are represented by single words. Two problems that may however frequently pose are: a) to distinguish homographs i.e. words that are spelt alike but having different meaning; b) words to be represented in their singular or plural forms.

Homographs are generally distinguished by providing the meaning by a scope word in brackets. For example: Tanks (Military)                  Tanks (Water)

Singular and plural forms are prescribed as given below with examples.

	Singular with examples	Plural with examples
i)	Abstract ideas; Literary forms E.g. Intelligence Density	Concrete objects or persons E.g. Airplanes Teachers
ii)	Biological species E.g. Palm Monkey	Larger groups E.g. Palms Monkeys
iii)	Names of fruits denoting fruit and the tree E.g. Peach Pear	Collections  E.g. Essay Biographies

**7.1132 Headings with two or more words (Phrase headings):** Frequently a subject is best expressed by a phrase.

**7.11321 Adjectival headings:** An adjectival heading consists of a noun modified by an adjective or an adjectival noun. The modifier may be a common adjective, a proper adjective, an ethnic, national, or geographic adjective, a common or proper noun in the possessive case, or a common or proper noun used as an adjective.

Ex: Computer architecture  
Social classes

**7.11322 Conjunctive phrase headings:** A conjunctive phrase heading consists of two or more nouns, with or without modifiers, joined by the conjunction "and." The purposes of this form of heading are (1) to express reciprocal relationships between two general topics discussed at a broad level from the perspectives of both topics, and (2) to combine concepts so similar that they are often treated together in documents.

Ex: Children and politics  
Boats and boating

**7.11323 Prepositional phrase headings:** A prepositional phrase heading consists of two or more nouns, with or without modifiers, linked by a preposition or prepositions. It is used to express complex relationships between topics which cannot be represented by a conjunctive phrase, or to represent a concept or thing only stated in the form of a prepositional phrase in the language. It is also used when there is an established pattern for similar headings.

Ex: Directors of corporations  
Doctor of philosophy degree

**7.11324 Inverted phrase headings:** Although words in headings are normally presented in their natural word order, certain phrase headings are inverted in order to bring the keyword, usually a noun, into the initial position. In a manual catalog or a single-entry listing or display, this inverted form serves to bring together subject headings containing the same noun for the purpose of subject collocation. Currently, natural word order is preferred in topical subject headings. The inverted form is retained in certain categories of headings involving cultural subjects and in cases where large numbers of similar headings in the inverted form have already been established.

Ex: Children's literature, Canadian  
Education, Higher  
Taxation, Exemption from

**7.11325 Free-floating phrase headings**

Certain phrases are designated as free-floating components which may be combined with any existing heading or with any heading within designated categories to form new phrase headings.

Ex: Johnson, Samuel, 1709-1784, in fiction, drama, poetry, etc.  
Manhattan (New York, N.Y.) in art  
Schools in literature

**7.114 Terminology:** In order to formulate a uniform heading for each subject, choices have been made among synonymous terms, between scientific and popular terms, among different language forms and variant spellings, and between current and obsolete terms.

**7.115 Cross-References:** Cross-references are made for the purposes of guiding the users from their entry vocabulary to valid headings and linking related headings. With 11<sup>th</sup> edition 1988 'See' and 'See also' references and the complementary xx and x were replaced by the thesaurus conventions BT, NT, RT, UF, USE and SA. Three types of relationships are represented in the cross-reference structure of Library of Congress Subject Headings: equivalence, hierarchical, and associative. These relationships are expressed in terms of USE, UF (Used for), BT (Broader term), NT (Narrower term), RT

(Related term), and SA (See also) references. Each reference links a term or heading with another heading or with a group of headings. For example:

**Ability – testing**

- SA subdivision ability testing under topical headings...
- UF Ability testing

**Ability testing**

- USE subdivision ability testing under subjects  
Ability –Testing

**Chemistry** [May Subd Geog]

- [QD]
- SA headings beginning with the word Chemical
- BT Physical sciences
- NT Acids  
Agricultural Chemistry

**Sun**

- SA headings beginning with the word solar

**7.116 Other features:**

**7.1161 Entry Format:** The formats of subject headings depend on their types. It makes use of various punctuation marks like comma, dash, hyphen, parentheses, square brackets, full stop, etc.

**7.1162 Notes:** Notes are provided under some headings in order to define the scope, to explain the relationships among headings, and to assist in the proper application of the headings so that consistency in assigning headings to documents on like subjects may be achieved. Relations to other headings and instructions, explanations, etc are given

**7.1163 Class Numbers:** A Library of Congress Classification number is added to a heading, if the caption for the number is identical or nearly identical in scope, meaning, and language to the subject heading, or if the topic is explicitly mentioned in an "Including" note under the caption for the number.

**7.1164 Filing Order:** LCSH's basic arrangement of subject headings is word by word. Numbers given in digits precede alphabetic characters in the order of increasing value. Initials separated by punctuation file as separate words. Abbreviations without interior punctuation file as single whole words.

- Ex: 4-H clubs  
A-36 (fighter bomber planes)  
ACI Test  
Alaska  
Children  
Children – Attitudes  
Children, Adopted

**7.117 Merits and demerits:**

The strongest aspect of LCSH is that it represents subject headings of a national library. The policy of the LCSH has made this as an undisputed leader in the field. Another strongest factor is it is revised every year. LCSH are also used as the indexing vocabulary in a number of published bibliographies. With all these strengths and support there are few problems with it. It is criticized for its outdated vocabulary, illogical syntax, general inefficiency for precise subject retrieval, the syndetic structure, and scope notes. LCSH meets the subject cataloguing requirements only for macro-documents

## **7.12 SEARS LIST OF SUBJECT HEADINGS**

Sears List of Subject Headings (SLSH) is an abridged version of the Library of Congress List of Subject Headings, named after the first compiler Minnie Earl Sears. It was in response to a demand from small libraries for a list of subject headings that Sears compiled this list to be less comprehensive and more suited with the needs of small public libraries and school libraries. Published in 1923, SLSH was initially based on the headings used by nine small libraries that were known to be well catalogued. Recognizing the need for uniformity, Sears brought out later the list adopting the principles and practices of LCSH, particularly to help those libraries that were using LC catalogue cards or wishing to add headings on the basis of LCSH. Presently it is in its eighteenth edition, (2004) published by the well known bibliographic publishers H W Wilson Company. SLSH is widely used today in and outside US by small libraries.

### **7.121 Scope of the Sears List:**

Sears List offers a basic list that includes many of the headings most likely to be needed in small libraries together with patterns and examples that will guide the cataloger in creating additional headings as needed. Headings for new topics can be developed from the Sears List in two ways, by establishing new terms as needed and by subdividing the headings already in the List. Instructions for creating new headings based on the pattern in Sears and sources for establishing the wording of new headings are given in the Principles of the Sears List.

### **7.122 Formation of Subject Headings**

Sears List of Subject Headings (SLSH) is also an enumerated list of subject headings. The structure of the headings is on similar patterns of LCSH, with some modifications to serve the needs of small public libraries and school libraries. SLSH is guided by general principles such as 'Specific and Direct Entry', 'Common Usage', 'Uniformity' in the formation of its subject headings. The rule of specific and direct entry is to enter the subject that accurately and precisely represents its contents. A document on PARROTS has to be entered under PARROTS, not under BIRDS or even COLOUR BIRDS. Another general rule is to use a popular or common name, rather than a scientific or technical name. for example, in a small public library, a reader may look for documents on BIRDS under BIRDS and not under ORINTHOLOGY. Another very important factor that governs the structure of headings in SLSH is synonyms. EARTHENWARE, CHINAWARE and PORCELAIN are all entered under PORCELAIN which must be applied consistently to all documents on this topic.

### 7.123 Forms of subject headings:

The simplest form of subject heading consists of a single noun, which is the most ideal type when it is possible to represent the thought contents of a document. The only snag that might cause ambiguity is only with homographs which are identical words with different meanings. Such words are identified by providing a scope word giving the contextual meaning of the word.

For example:           SEALS (Animals)  
                          SEALS (Numismatics)

Choosing between singular and plural forms of single words, the prescription is as follows: Abstract ideas, concept or action are usually stated in the singular, whereas objects and things are rendered in their plural forms.

For example:           HEALTH                   (Singular)  
                          EMPLOYEES               (Plural)

With regard to two worded single concepts i.e. nouns qualified by adjectives SLSH recommendation is to stress the key word to avoid scattering material on the same subject throughout the alphabet. The noun is placed first in order to keep all aspects of a broad subject together when that result is deemed desirable. Inversion can be made when the first element qualifies the second and the second is an independent unit. For example:

Art, Abstract	Education, Elementary	Insurance, Accident
Art, Chinese	Education, Higher	Insurance, Fire
Art, Municipal	Education, Secondary	Insurance, Health

It should, however, be noted that some adjective noun phrases could never be inverted because the noun has no significance without the adjective, e.g. International relations.

#### 7.1231 Complex Subjects

Dealing with complex subject i.e. documents that deal with more than one concept in a particular relation, the main and sub-divisions are stated together.

For example:   Birds – Eggs and nests           Earth – Age  
                  Birds – Migration               Earth – Crust  
                  Birds – Protection           Earth – Internal structure

Examples of form headings:   Artists – Directories  
  Children's literature – Bibliography  
  Economics – Dictionary

Examples of headings presented from a particular point of view:

Education – History  
Mathematics – Philosophy  
Psychology – Research  
Sociology – Study and Teaching

Examples of geographic names: Subject divided by place:

Agriculture – India  
Theater – Paris

Examples of geographic names: Names of places divided by subject:

Karnataka – History  
China – Boundaries  
Tibet – Climate

In summary, the formation of subject headings in SLSH is a little more or less complicated and straightforward, but based generally on the practices of LCSH.

### 7.124 Entry Format

Subject entries in SLSH are printed in boldface while See references appear in light faces. In the list, the right half of each page is blank. All entries, references and instructions are confined to the left columns, to leave space for the local cataloguer to add any new headings. References or comments needed to convert the volume into an authority file.

The general format of entry in SLSH is as follows:

Subject Entry in boldface	<b>LABOUR UNIONS</b>
Scope Note	(in this case, there is no scope note)
See also references in italics	<i>See also</i> Arbitration, Industrial; Collective bargaining; Injunctions; Open and closed shop; Sabotage; Strikes and lockouts; Syndicalism; also names of types of unions and names of individual labour unions, e.g. Librarians' unions; United Steelworkers of America; etc.

*See* references indicated by x in italics      x Labour organizations; Organised Labour; Trade Unions; Unions, Labour

Reversible *See* also indicated by xx in italics

xx Collective bargaining; Cooperation; Industrial relations; Labour; Socialism; Societies; Strikes and lock outs

### 7.125 Cross References

There are three types of cross references viz. Specific 'See' references; Specific 'See also' references; and General references.

The See references are concerned mainly with terminology, guiding the reader from words the person may think of to those actually used for subject headings. See references are generally made from:

- Synonyms and near synonyms; e.g. Lifts see Elevators
- Second part of a two worded heading; e.g. Technical chemistry see CHEMISTRY Technical.
- Conjunctive i.e. terms connected by and; e.g. Crime and Narcotics see NARCOTICS AND CRIME
- Inverted headings to normal order; e.g. Education, Adult see ADULT EDUCATION
- Variant spellings; e.g. Colour see Color
- Opposites; Intemperence see TEMPERENCE
- Singular to plural; e.g. Mouse see MICE

See also references are concerned entirely with guiding the reader from headings where he has found some information to other headings that list materials on related or more specific aspects of the subject of his enquiry.

For examples: HEALTH (N.B. Note the variety of aspects of the subject)

See also

Diet	Longevity
Disease	Mental health

Exercise	Physical fitness
Health education	Rest
Hygiene	Sleep

### **7.126 Salient Features of 18<sup>th</sup> Edition**

There are three major features of this new edition of the Sears List. The first is the inclusion of five hundred new subject headings. The second is the revision of the classification numbers to conform to the usage of the 14th edition of the Abridged Dewey Decimal Classification (2004). The third is a small but important addition to the Principles of the Sears List. It has been expanded in this edition to provide guidance to libraries that choose to assign topical and geographic headings to individual works of fiction, drama, and poetry. The List of Commonly Used Subdivisions, which was omitted in the previous edition of the Sears in favor of a more exhaustive treatment of subdivisions within the body of the List, has been restored in this edition and renamed List of Subdivisions Provided for in the Sears List. It now lists, for the purpose of easy reference, every subdivision for which there is a provision in Sears, no matter how specialized. At the same time, for every subdivision there is an entry in the alphabetical List with full instructions for the use of that particular subdivision.

Many of the headings new to this edition were suggested by librarians representing various sizes and types of libraries, by commercial vendors of bibliographic records, and by the catalogers, indexers, and subject specialists at the H.W. Wilson Company. The most significant is the replacement of the subdivision Description by Description and travel. Certain headings of decreasing interest and some unnecessary examples, such as Margarine, Van life, and Iran-Contra Affair, 1985-1990, have been deleted from the List. Such headings are not invalid and may be maintained in the catalog.

It was the policy of Minnie Sears to use the Library of Congress form of subject headings with some modification, chiefly the simplification of phrasing. The Sears List still reflects the usage of the Library of Congress. A major difference between the two lists is that in Sears the direct form of entry has replaced the inverted form, on the theory that most library users search for multiple-word terms in the order in which they occur naturally in the language. In most cases cross-references have been made from the inverted form.

As in previous editions, all the new and revised headings in this edition have been provided with scope notes where such notes are required. There are also scope notes in Sears that identify any headings in the area of literature that may be assigned to individual works of drama, fiction, poetry, etc.

The classification numbers in this edition of Sears are taken from the 14th edition of the Abridged Dewey Decimal Classification (2004). For spelling and definitions the editor has relied upon Webster's Third New International Dictionary of the English Language, Unabridged (1961) and the Random House Webster's Unabridged Dictionary, 2nd ed., revised and updated (1997). Capitalization and the forms of corporate and geographic names used as examples are based on the Anglo-American Cataloguing Rules, 2nd ed., 2002 revision. The filing of entries follows the ALA Filing Rules (1980). Every term in

the List that may be used as a subject heading is printed in boldface type whether it is a main term; a term in a USE reference; a broader, narrower, or related term; or an example in a scope note or general reference. If a term is not printed in boldface type, it is not used as a heading.

### **7.127 Merits and Demerits:**

For 80 years, Sears List of Subject Headings has served the needs of small and medium-sized libraries, delivering a basic list of essential headings, together with patterns and examples to guide the cataloger in creating further headings as needed. Practical features include a thesaurus-like format, an accompanying list of cancelled and replacement headings, and legends within the list that identify earlier forms of headings. It is comparatively simple to use than the Library of Congress List of subject Headings. The rules and principles are fairly explicit in their directions, containing scope notes and specific instructions for their use. The major inherent problem with SLSH is that it is not backed by any library collection and as such updating and revision of subject headings cannot keep pace with changing current terminology. Another comment on this is that it does not have its own theoretical foundations. As the subject headings have been drawn from the vast collections of the Library of Congress in a wide variety of subjects, it lends itself an authority as a subject cataloguing reference tool, particularly macro level documents. An important reason for its widespread use is the fact the fact that the Library of Congress cataloguing records have been available to other libraries through MARC records.

## **7.2 THESAURUS:**

### **7.21 DEFINITION:**

A thesaurus is a tool for vocabulary control. Usually, a thesaurus is designed for indexing and searching in a specific subject area. (You will study MeSH and INSPEC Thesaurus in the section 7.24). The word thesaurus is derived from 16th century New Latin, in turn from Latin *thesaurus*, from ancient Greek *θησαυρός* *thesauros*, "store-house", "treasury". Besides its meaning as a treasury or storehouse, it more commonly means a listing of words with similar, related, or opposite meanings (this new meaning of *thesaurus* dates back to *Roget's Thesaurus*). *Roget's Thesaurus*, was published in 1852, having been compiled earlier, in 1805, by Peter Roget. Entries in *Roget's Thesaurus* are not listed alphabetically but conceptually and are a great resource for writers.

In Information Technology, a thesaurus represents a database or list of semantically orthogonal topical search keys. In the field of Artificial Intelligence, a thesaurus may sometimes be referred to as an ontology. A formal definition of a thesaurus designed for indexing is: a list of every important term (single-word or multi-word) in a given domain of knowledge; and a set of related terms for each term in the list. Terms are the basic semantic units for conveying concepts. They are usually single-word nouns, since nouns are the most concrete part of speech. Verbs can be converted to nouns -- cleans to cleaning, reads to reading, and so on. Adjectives and adverbs, however, seldom convey any meaning useful for indexing. When a term is ambiguous, a "scope note" can be added to ensure consistency, and give direction on how to interpret the term.

## 7.22 WHAT IS IN A THESAURUS?

A thesaurus gives several types of information to indexers and searchers.

**Preferred Terms:** The thesaurus indicates which terms' indexers and searchers are allowed to use. These terms are called preferred terms. This is a major part of vocabulary control – restricting the vocabulary so that it is easier to predict what words might have been used to index a concept.

**Non-preferred Terms:** In addition to preferred terms, a thesaurus also needs to indicate some terms that indexers and searchers are not to use. These terms are called non-preferred terms. It should be possible to look up a non-preferred term and see what preferred term should be used instead. This will save time and make it less likely that the best preferred term will be missed. A thesaurus also usually allows to look up a preferred term and see its non-preferred terms. This can give you a better idea of what the term is supposed to mean.

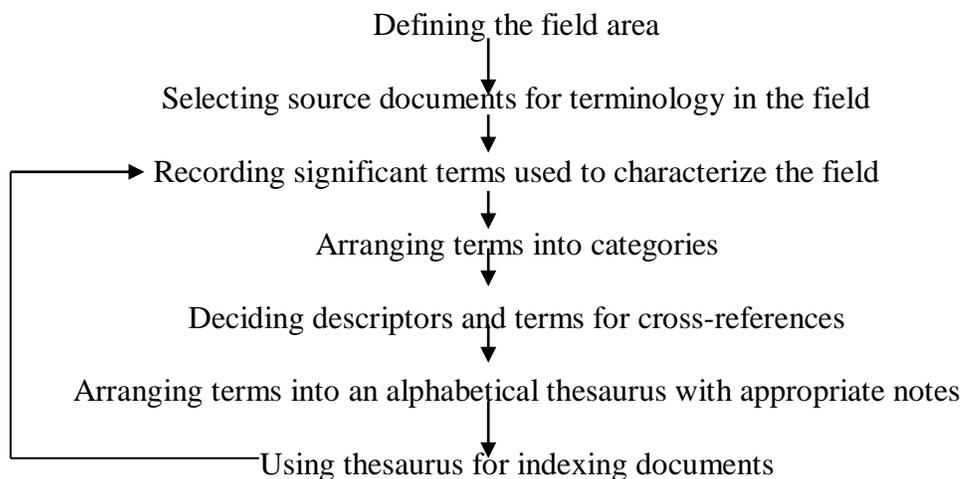
**Semantic Relations:** As well as linking preferred terms with non-preferred terms, a thesaurus also shows links between different preferred terms. These links are usually for semantic relations. Like a link between a preferred term and a non-preferred term, one of these semantic links can help to direct you to the right term and make the meaning of a term clearer.

**Guides to Application:** A good thesaurus should make it clear what a term is meant to cover. It can accomplish this to some extent by showing non-preferred terms and semantic relations. Other ways of guiding people in using a thesaurus include introductory matter and scope notes. A scope note often takes the form of a definition of the term. Ensuring that terms are used consistently with the same meaning is another major aspect of vocabulary control.

**Rules for Synthesis:** Usually, a thesaurus lists all its preferred terms explicitly. Such thesauri are enumerative. Some thesauri indicate some preferred terms indirectly: instead of listing all the preferred terms, they give rules for creating them out of components. Such thesauri are at least partly synthetic.

## 7.23 CONSTRUCTION OF THESAURUS:

The general steps to be taken in thesaurus construction are given below



### 7.231 Collecting Terms

Thesaurus construction requires collecting a set of terms. Some of these will end up becoming preferred terms and others may not appear in the thesaurus at all in their original form, but they may suggest concepts that need to be covered in some way. Sources from which terms can be collected include

- Existing lists of terms, other thesauri, indexes, dictionaries, glossaries, etc.
- Texts from which terms can be extracted: titles, abstracts, or full texts of indexed items, queries by patrons
- People: subject specialists, etc.

#### 7.2311 What Kinds of Terms should be collected?

Where possible, terms in a thesaurus should be nouns or noun phrases. A term should be general enough that it might be used to index a number of items. For example, a thesaurus usually does not include proper names. But a term should not be so general that it might be used to index too many of the items in the thesaurus' subject area. For example, the term "NEWS" would not be much use in a thesaurus for indexing news items.

### 7.232 Modifying and Inventing Terms

#### 7.2321 Standardizing the Form of Words

Terms collected should already be nouns or noun phrases. Here are some further guidelines for the form that terms should take in your final thesaurus.

Guidelines	Examples
Plural for things that can be counted	"TUBES"
Singular for "mass" nouns	"WOOD"
Singular for processes, properties, and conditions	"REFRIGERATION" "WEIGHT", "POVERTY"
Not inverted	"RADAR ANTENNAS" (rather than "ANTENNAS, RADAR")
Excluding prepositions	"CARBOHYDRATE METABOLISM" (rather than "METABOLISM OF CARBOHYDRATES")
Excluding punctuation marks, diacritics, special characters, and abbreviations	"COOPERATIVE PROGRAMS" (rather than "CO-OPERATIVE PROGRAMS" or "COÖPERATIVE PROGRAMS") "MUSICAL NOTES" (rather than "(MUSICAL) NOTES" or "MUS. NOTES")

**7.23211 What to do with terms with more than one meaning:** A homograph is an expression that has the same spelling as another expression, but a different meaning. A thesaurus needs to distinguish between homographs. A unique term may be created out of

a homograph by adding a parenthetical qualifier; for example, "PORT (WINE)". You may note that including parentheses is contrary to the guideline given above; namely, to avoid punctuation. A unique term may also be created out of a homograph by adding another word without punctuation; for example, "PORT WINE".

**7.23212 Introducing New Terms:** In addition to terms extracted from your various sources, you may sometimes choose to introduce new terms of your own.

**Broad Concept Terms:** Terms that represent broad concepts may be introduced because they are useful in broad searches. For example, "TRAFFIC STATIONS", because it can be used to replace a search for "AIRPORTS OR BUS TERMINALS OR TRAIN STATIONS OR HELIPORTS OR...".

**Structural Terms:** Terms may also be introduced because they help to clarify the structure of semantic relations. For example, "EMPLOYMENT OF SPECIFIC GROUPS" to clarify the relationship between "EMPLOYMENT" and "YOUTH EMPLOYMENT".

**Terms for Non-textual Material:** If you are constructing a thesaurus for indexing material, which is not in the form of text, you have fewer sources for terms. You may therefore find yourself inventing your own terms more.

### **7.233 Preferred Terms and Non-preferred Terms**

**7.2331 Equivalent Terms:** After collecting terms for your thesaurus, you need to decide which equivalent terms are. For purposes of indexing and searching, a set of equivalent terms will all be treated as though they meant the same thing and will be represented by a single preferred term.

**7.2332 Spelling and Synonyms:** Sometimes, equivalent terms really do mean the same thing. So, it obviously makes sense to use a single preferred term to represent that one meaning.

A word may have more than one spelling; for example, "AESTHETICS" and "ESTHETICS". Two different words may have essentially the same meaning; for example, "AUTOMATION" and "MECHANIZATION".

**7.2333 Quasi-synonyms:** Sometimes, equivalent terms mean different things in ordinary language. For indexing and retrieval, it is better to group the different meanings together. Such equivalent terms are called quasi-synonyms.

Ex: 1. Terms with overlapping meanings, "GENIUSES" and "PRODIGIES" might be treated as equivalent, even though the two terms mean different things.

2. A term whose scope is included in that of another term, "STEEL" might be treated as equivalent to "METAL" if it is not important to distinguish items on steel from items on other metals.

3. Sometimes opposites are treated as equivalent, because items on one are likely to be relevant to a query for the other. For example, "TRANSPARENCY" might be treated as equivalent to "OPACITY".

**7.2334 Preferred Terms:** Preferred terms serve as focal points where all the information about a concept is collected.

**7.2335 Non-preferred Terms:** Non-preferred terms are included in a thesaurus mainly to help users find the appropriate preferred terms. Non-preferred terms may also help to define the scope of preferred terms.

**7.2336 USE/UF:** A non-preferred term is normally linked to a corresponding preferred term by a USE reference. The corresponding reference in the opposite direction is UF ("Used For").

For example,

PERIODICALS	SERIALS
USE SERIALS	UF PERIODICALS

Here the preferred term is "SERIALS" and the corresponding non-preferred term is "PERIODICALS".

**7.2337 Choosing Preferred Terms:** The following are some principles for choosing preferred terms, together with examples of applying them.

Guidelines	Examples
Usage	COOKING UF COOKERY ("Cooking" is the more commonly used word.)
Breadth	PLASTICS UF POLYETHYLENE ("Plastics" clearly means all plastics, of which polyethylene is only one.)
Disambiguation	AMERICAN LIBRARY ASSOCIATION UF ALA ("ALA" could stand for something else.)
Collocation	RAILWAY STATIONS UF TRAIN STATIONS (In an alphabetical sequence, "RAILWAY STATIONS" would appear near to "RAILWAYS" and other terms related to railways.)
Conciseness	MUCKRAKERS UF MUCKRAKING MOVEMENT (One word rather than two.)
Plural for countable objects	GEESE UF GOOSE (Geese are countable.)
Internal consistency	If you have decided to prefer the Latin names for plants, do so consistently.
External consistency	You might prefer "PIERS & WHARVES" to "LANDINGS", "BOAT LANDINGS", "DOCKS", "QUAYS", or "WHARVES" partly because that is what the Library of Congress Thesaurus for Graphic Material does.

**7.2338 Compound USE References:** Instead of a single non-preferred term, one may sometimes instruct indexers and searchers to use more than one preferred term in combination. In such cases, the USE reference points to all the preferred terms, and the UF reference is often marked in some special way.

For example,

SNOWMOBILES USE VEHICLES+SNOW
SNOW UF+ SNOWMOBILES
VEHICLES UF+ SNOWMOBILES

You are especially likely to do this if the non-preferred term consists of more than one word.

For example,

SCHOOL CAFETERIAS USE CAFETERIAS+SCHOOLS
CAFETERIAS UF+ SCHOOL CAFETERIAS
SCHOOLS UF+ SCHOOL CAFETERIAS

On the other hand, you may choose not to make such a term a non-preferred term, even if it consists of more than one word.

**7.2339 Making Multi-word Terms Preferred:** A term consisting of more than one word should typically be made a preferred term if combining terms is not possible either at the indexing stage or at the searching stage. Too many terms would otherwise be required to index an item or the resulting number of preferred terms is not too large. Indexing and searching are generally easier using the compound term. The term is likely to be used frequently in indexing or searching. The term's components occur frequently in different syntactic relations; for example, "LIBRARY SCHOOLS", "SCHOOL LIBRARIES". The term is needed in the structure of semantic relations; especially, if any narrower concepts are represented by preferred terms.

### 7.234 Semantic Relations

**7.2341 Why Indicate Semantic Relations?** Indicating semantic relations helps in several aspects of information management:

- checking whether a term should be used in indexing a given item or in formulating a given search specification
- choosing the correct level of generality in indexing and searching
- searching in response to broad inclusive queries
- sharing indexing by facilitating translation from one scheme to another

**7.2342 Semantic Relations between Terms:** The main semantic relations indicated between preferred terms in a thesaurus are hierarchical relations and non-hierarchical relations.

**7.23421 BT and NT Links:** BT and NT links are used to indicate hierarchical relations. In a hierarchical relation, one term is viewed as being "above" another term because it is broader in scope. In developing a thesaurus, it is often a good idea to work out the hierarchical relations first.

**7.23422 When is there a Broader/Narrower Term Relation?** There are various definitions of what constitutes a hierarchical relation. In this Unit the following relations are explained.

**Genus/Species:** Assume Term A is a broader term to term B (and term B is a narrower term to term A) if all the things included in the class named by term B are included in the class named by term A. For example, "ANIMALS" is a broader term to "CATS" (and "CATS" is a narrower term to "ANIMALS") because all cats are animals. On the other hand, "PETS" is not a broader term to "CATS" because not all cats are pets.

**Class/Member:** The narrower term can sometimes name a class with only one member. For example, "UNIVERSITIES" is a broader term to "KARNATAKA STATE OPEN UNIVERSITY, MYSORE " because KARNATAKA STATE OPEN UNIVERSITY is a university. Since thesauri usually do not include proper names, you may not encounter cases like this in constructing your own thesaurus.

**Hierarchical Whole-Part:** Term A is a broader term to term B (and term B is a narrower term to term A) if everything included in the class named by term B is a part of something included in the class named by term A. For example, in a medical thesaurus, "HEAD" might be a broader term to "NOSE" because noses are normally parts of heads. On the other hand, "FORESTS" would not be a broader term to "TREES" because not every tree is part of a forest.

**Geographical Whole/Part:** In a hierarchical whole/part relation, both the broader term and the narrower term may name a class with only one member. This is often true of geographical names. For example, "SOUTH INDIA" is a broader term to "KARNATAKA" because 'Karnataka' is a part of 'South India'. On the other hand, "KARNATAKA" is not a broader term to "CAUVERY RIVER" because only part of the river is part of Karnataka. Since many thesauri do not include geographical names, you may not encounter cases like this in constructing your own thesaurus.

### **7.235 BT, NT, and RT References**

**7.2351 What is the relationship between BT and NT?** Normally, BT and NT are "inverse" links. In other words, if X is a broader term to Y, then Y is a narrower term to X, and vice versa. For example, if a thesaurus contains the entry

PENS  
BT WRITING MATERIALS

You would expect it also to have the entry

WRITING MATERIALS  
NT PENS

**7.2352 How many BT references can a term have?** A thesaurus is usually "poly hierarchical". This means that a term can have more than one immediately broader term and more than one BT reference. For example,

SOCIAL PSYCHOLOGY  
 BT PSYCHOLOGY  
 BT SOCIOLOGY

Poly- hierarchy avoids futile arguments about the "best" broader term to choose. Some terms in a thesaurus have no broader terms and so no BT references. Such terms are usually fairly broad in meaning, at least within the subject area covered by the thesaurus. For example, in a sports thesaurus, "SPORTS" might have no broader terms.

**7.2353 When should BT/NT references be omitted?** You should not indicate every hierarchical relation explicitly in your thesaurus. That could make the entries too long and difficult to read. Instead, omit those links that are implied by other links. Suppose X is a broader term to Y, which in turn is a broader term to Z. Do not make BT/NT references between X and Z. For example,

PLANT PRODUCTS  
 NT FRUIT

and

FRUIT  
 NT FRESH FRUIT

but not

PLANT PRODUCTS  
 NT FRESH FRUIT

**7.2354 When to Use an RT Reference?** An RT reference is used for non-hierarchical semantic relations in a thesaurus. To decide whether there should be an RT reference between two preferred terms X and Y that do not have a hierarchical relation; you can use the following test:

- Should an indexer/ a searcher considering using X be reminded of the existence of Y?
- What Is the Relationship between RT and RT?
- Normally, RT is its own "inverse" link type. In other words, if X has an RT reference to Y, then Y should have an RT reference to X. For example, if a thesaurus contains the entry

PENS  
 RT CALLIGRAPHY

you would expect it also to have the entry

CALLIGRAPHY  
 RT PENS

**7.2355 Semantic Categories of RT References:** In constructing your thesaurus, you may find it useful to list some categories of semantic relations that you think should be covered by RT references. Here are some categories sometimes used, with examples.

Categories	Examples
Time	LEISURE READING

	RT LEISURE TIME
Place	FOREIGN LANGUAGES RT LANGUAGE LABORATORIES
Product	STILL CAMERAS RT PHOTOGRAPHS
	SHIPBUILDING RT SHIPS
Cause	VANDALISM RT HOSTILITY
Agent	COACHING RT COACHES
Device	PAINTING RT PAINT BRUSHES
Application	COMPUTERS RT WORD PROCESSING
Part	VEHICLES RT WHEELS
Complement	PARENTS RT CHILDREN

**7.236 Scope Notes:** The most common type of guide to applying terms in a thesaurus is the scope note. A scope note is normally preceded by the notation SN. Scope notes take a variety of forms. Scope notes may give definitions, indicate which concepts are included or excluded, refer to other terms, provide additional instructions and they should be relevant to indexing and searching well-formed. Information included in a scope note should be helpful to users of the thesaurus as indexers or searchers. It should add to what the term already says by itself. Simply repeating the term or giving an obvious definition of an unambiguous term is not helpful. Remember that a thesaurus is not a dictionary, an encyclopedia, or even an index. Scope notes should be well formed. They should contain no spelling errors. Many scope notes do not use complete sentences. You can use noun and verb phrases instead. Nevertheless, the syntax should be correct. For example

SPACE ERROR

SN TENDENCY TO BE BIASED BY THE SPATIAL POSITION OF STIMULI IN  
RELATION TO THE OBSERVER

BEARS

SN DOES NOT INCLUDE PANDAS

**7.237 Thesaurus Displays:** For any thesaurus display, you may need to make several decisions. These decisions are likely to include: which types of terms will have entries, how to indicate special types of terms, what types of links will be shown to other terms, how many levels of linking will be shown, how to indicate link types, where the linked terms are placed, relative to the entry term, relative to each other. Any of these decisions

will, of course, be constrained in various ways. For example, the thesaurus construction software that you use may produce only certain kinds of displays or may not permit you to store a mixture of upper and lower case.

**7.2371 Which types of terms will have entries?** A thesaurus display might have entries only for preferred terms. At least one of the displays, however, should provide entries for non-preferred terms as well, to allow users to browse through these for the correct preferred terms. One of the displays might include entries only for top terms, preferred terms that have no broader terms. This choice is often combined with indicating multiple levels of narrower terms, in a tree display, as discussed below.

...  
EX-CONVICTS  
EYE EXAMINATIONS  
EYE PATCHES  
EYES  
FABLES  
FABRIC DESIGN DRAWINGS  
...

**7.2372 How to indicate special types of terms?** You may want to mark certain kinds of terms in special ways. For example, you might put all the non-preferred terms in italics:

...  
*EXTREMISM*  
EYEGLASSES  
EYES  
FABLES  
*FABRIC DESIGN DRAWINGS*  
...

Of course, users of the thesaurus should be able to tell that a term is a non-preferred term if it has a USE reference after it, but displaying the term differently will serve as an added reminder. The examples used in this Unit generally show terms in all upper case. This is to emphasize that the distinction between upper and lower case should normally not be significant in indexing and searching using a controlled vocabulary. Nevertheless, you may prefer a mixture of upper and lower case for your thesaurus displays to make them easier to read. Mixing upper and lower case may be especially helpful for longer elements such as scope notes:

GOGGLES  
SN *Protective eye coverings.*

**7.2373 What types of Links should be shown to other terms?** Taken as a whole, your thesaurus displays should cover all the term links that are important to the people who will use the thesaurus. In individual displays, you may choose to include only certain links. For example, "USE" references only or all the links or the scope notes have to be displayed,

ALIDADES  
SN *TELESCOPIC SITING DEVICES USED AS PART OF A SHIP'S NAVIGATIONAL*

*EQUIPMENT FOR TAKING BEARINGS*  
BT SCIENTIFIC EQUIPMENT  
BT TELESCOPES  
RT NAVIGATION

**7.2374 How many levels of linking will be shown?** In one of your displays, you may wish to show indirectly linked terms as well as those linked directly to the entry term. This is mostly useful with links representing hierarchical relations. The display could indicate more than one level of broader term:

MONOCLES  
BT EYEGASSES  
. BT OPTICAL DEVICES  
. . BT EQUIPMENT  
. BT MEDICAL EQUIPMENT AND SUPPLIES

**7.2375 How to indicate link types?** You can often omit the symbols for the different kinds of links if it is obvious what they are. So, you need to use a symbol such as "RT" only once for a series of terms all linked to the entry term with an "RT" reference:

...  
RT EMPLOYEE EATING FACILITIES  
    EMPLOYEE FRINGE BENEFITS  
    EMPLOYEE RIGHTS  
    EMPLOYMENT  
    LABORERS  
    UNEMPLOYED

Similarly, if all the links in a display are of the same kind, as in a hierarchical display, you do not need to use a distinctive symbol:

EQUIPMENT  
. AIRPLANE EQUIPMENT  
. . AIRPLANE PROPELLERS  
. . AIRPLANE WINGS  
. ANCHORS  
. APPLIANCES  
. . AIR CONDITIONERS  
. . DISHWASHING MACHINES  
...

**7.2376 Where the linked terms are placed:**

**7.23761 Relative to the Entry Term.** In a printed display, you will usually want the entry term to appear at the top left, because this makes it easy to search for. Variations are possible, though. For example, broader terms may be displayed above the entry term and narrower terms below:

. MEDICAL EQUIPMENT AND SUPPLIES  
 .. EQUIPMENT

In an online display of a single entry, there is more flexibility. For example, the broader terms can be arrayed on the left, the narrower terms on the right, and the related terms above and below:

MEDICAL EQUIPMENT AND SUPPLIES . EQUIPMENT . . OPTICAL DEVICES .	CONTACT LENSES	. MONOCLES . SUNGLASSES
	EYE PATCHES	
	EYEGLASSES	
	GOGGLES	

**7.23762 Relative to each other:** If all the links are indicated in the same position relative to the entry term, the best order to follow is generally scope notes, non-preferred equivalent terms, broader terms, narrower terms, related terms. For example,

EMPLOYEES  
 SN *PERSONS IDENTIFIED AS WORKING FOR ANOTHER, BUT WHERE THE NATURE OF THE OCCUPATION, BUSINESS, OR INDUSTRY IS NOT KNOWN.*  
 UF *PERSONNEL*  
     *STAFF*  
     *WORKERS*  
 BT PEOPLE  
 NT HOTEL EMPLOYEES  
     RAILROAD EMPLOYEES  
 RT EMPLOYEE EATING FACILITIES  
     EMPLOYEE FRINGE BENEFITS  
     EMPLOYEE RIGHTS  
     EMPLOYMENT  
     LABORERS  
     UNEMPLOYED

Within a group of terms linked in the same way to the entry term, the order is most commonly alphabetical, as in the example just given. Sometimes, however, you may wish to adopt a systematic order by subcategorizing the link types, especially if the lists are very long.

**7.24 STUDY OF THESAURUS:** In this section you will be introduced to two important thesauri: 1. Medical Subject Headings and 2. INSPEC Thesaurus.

**7.241 MEDICAL SUBJECT HEADINGS (MeSH®)**

MeSH (Medical Subject Heading) (<http://www.nlm.nih.gov/mesh/meshhome.html>) is the National Library of Medicine USA's controlled vocabulary thesaurus in operation since 1964 developed to provide suitable headings for both the current catalog and IM of NLM. The present online file dates from 1966. MeSH now appears in two formats – annual publication and annotated alphabetic list intended for online users.

It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. MeSH descriptors are arranged in both an alphabetic and a hierarchical structure. At the most general level of the hierarchical structure is very broad headings such as "Anatomy" or "Mental Disorders." More specific headings are found at more narrow levels of the eleven-level hierarchy, such as "Ankle" and "Conduct Disorder." There are 22,997 descriptors in MeSH. In addition to these headings, there are more than 151,000 headings called Supplementary Concept Records (formerly Supplementary Chemical Records) within a separate thesaurus. There are also thousands of cross-references that assist in finding the most appropriate MeSH Heading, for example, Vitamin C *see* Ascorbic Acid. These additional entries include 24,050 printed *see* references and 112,012 other entry points.

The MeSH thesaurus is used by NLM for indexing articles from 4,800 of the world's leading biomedical journals for the MEDLINE/PubMED® database. It is also used for the NLM-produced database that includes cataloging of books, documents, and audiovisuals acquired by the Library. Each bibliographic reference is associated with a set of MeSH terms that describe the content of the item. Similarly, search queries use MeSH vocabulary to find items on a desired topic.

**7.2411 Subheadings:** There are four types of subheadings: topical, form, language and geographic. Topical subheadings are for use by cataloguers, indexers and searchers.

**Publication types:** These were introduced in 1991 to specify the nature of the information or the way in which it is published, to expand on the previous citation types; they do not relate to the subject of the document. There are more than 50 publication types in the list.

**7.2412 Check Tags:** There are check tags which were introduced after the 1965 review to specify the kind of study indexed. For example, it is important to know whether a trial of a new drug is on animals or people, in the field or in the laboratory, so we find check tags like ANIMAL, HUMAN etc.

**7.2413 Geographics:** For indexing geographic headings are taken from Category Z of the tree structures.

**7.2414 Non-MeSH headings:** Some headings appear in the tree structures as necessary steps of division but are not helpful indexing or searching terms. For example, UNITED STATES BY INDIVIDUAL STATE is an essential part of the tree structure, but is hardly likely to be used as an indexing term or sought by user. These terms are referred to as Non-MeSH terms and can serve a very useful purpose in searching.

**7.2415 Cross-references:** MeSH does not use the conventional BT,NT, RT codes for linkages. Instead it has a very explicit set developed particularly for MEDLARS. Since 1991, 'consider also', 'see' and 'see related' links have been introduced to draw the user's attention to terms which may be linguistically related.

Ex:   Pregnancy     See related   Pre natal care  
      Brain/injuries see Brain injuries  
      Kidney consider also Glomerul

**7.2416 Notes:** The notes, which are found under a high proportion of headings, are equally helpful. These too are of various kinds like: right word?, definition, use notes for searching, permitted subheadings, history of the term, and annotations etc.

**7.2417 Revision:** The list is updated annually, with particular attention paid to areas which appear to need revision, as well as to the addition of new terms.

**7.2418 Tree structure:** The alphabetical lists are complemented by the categorized list containing the tree structures. These form a detailed classification of all the concepts in MeSH, including the Non-MeSH headings. There are 15 major categories denoted by the letter A to N and Z (the geographic listing).

**7.2419 Comments:** MeSH is a very powerful tool. There is a great deal of help for the indexer and the list is under constant review to keep it up to date. While it is too closely geared to the field of medicine to act as a direct model, there are many features which other thesauri might well learn from and a great deal of effort has obviously been expended to make sure that there will be as much consistency as possible between indexers and searchers.

### **7.242 INSPEC THESAURUS:**

Produced by the Institution of Electrical Engineers (IEE), INSPEC is the world's largest English-language bibliographic database in the fields of physics, electrical engineering and electronics, computers and control engineering and information technology. This was first published in 1973 and is revised every two years. The INSPEC Thesaurus 2004 has undergone an extensive revision to incorporate INSPEC's more comprehensive coverage of manufacturing and production. The INSPEC Thesaurus is available as part of the INSPEC Search Aids CD-ROM, which combines the INSPEC Thesaurus, INSPEC Classification and List of Journals on one CD. Other Electronic versions of the INSPEC Thesaurus and the User Documentation Package are also available. A print version is available for the INSPEC Thesaurus 2004.

The thesaurus is in two parts. The first and the major part is the alphabetical list of terms, while the second is the list of term trees. The Thesaurus now contains about 9,000 preferred terms of which 673 are new. Each INSPEC record is indexed using controlled terms chosen from the Thesaurus to provide a powerful search aid. As well as listing the controlled terms and lead-ins or cross reference terms, the Thesaurus give further help by showing the relationships between terms, the dates on which they were added and the terms in use before these dates. All terms are in lower case, except proper nouns. Preferred nouns are in bold, with non-preferred terms in normal type, e.g:

#### **Thermal insulation**

UF     heat insulation  
          insulation, thermal

heat insulation

USE    thermal insulation

In cases where a term is used instead of the inverted form, e.g. insulation, thermal in the above example, a NT reference from the term qualified replaces the USE reference.

Insulation

NT thermal insulation

A preferred term may be used for more than one non-preferred term. Some terms have scope notes. These are not designated SN, but appear in italics immediately after the term. All preferred terms have date inputs. For terms introduced or changed since that date the dates previously used are listed as Prior Terms (PT). Class Numbers from the INSPEC classification are given (CC). One term may have more than one class number, depending on context.

Aberrations [broad term used for several specific terms]

*Aberrations in optics and particle optics only*

UF astigmatism (optical)  
Sedel theory [Proper noun]  
RT aspherical optics  
CC A4180; A420F; A4278  
DI January 1969

#### **7.2421 Hierarchical Display:**

As well as the usual BT, NT, and RT links the thesaurus gives TT for 'Top Term' in the hierarchy.

#### **Cable television**

UF CATV  
BT television  
TT telecommunication

The hierarchies are shown in the second part of the list in alphabetical order of top terms; since there may be more than one hierarchy linking terms, a given term may appear more than once in the display, either in more than one hierarchy or more than once within the same top term.

Alloys

- . Germanium alloys
- .. Ge-Si alloys
- . Silicon alloys
- .. Elinbvar

**7.2422 Comments:** It is not too difficult to find points of criticism in any thesaurus or classification scheme, but this should not hide the fact that the INSPEC thesaurus and classification together provide a powerful tool for information retrieval. The headings in the classification are used to arrange the printed versions, but can also be used in searching the database. The thesaurus is updated quite regularly to provide users with the most effective tool possible. The omission of qualifying terms does keep the size of the index down, but at a price. To check every one of the eight places listed for, e.g. stimulated emissions or strain gauges would surely become tedious.

### 7.3 THESAUROFACET:

The English Electric Company had a very active library service and the librarian and other members of the CRG devised the EE Classification of Engineering, which reached its third edition in 1961. As the library began to use computer techniques and post coordinate indexing, it was decided that a new version should combine a classification and a thesaurus. This resulted in the designing of 'Thesaurofacet' in 1969 by Jean Aitchison. The thesaurofacet was the integration of classification schedules and thesaurus. The full title "Thesaurofacet: A thesaurus and faceted classification for engineering and related subjects" indicates that there are two tools here, a classification and a thesaurus, but it is worth to note that the two have to be used if the best results are to be derived.

As has been pointed out earlier elsewhere, a classification scheme can only display one set of genus-species relationships at a time, though a particular term may appear in more than one hierarchy, of course. The alphabetical sequence was used as the index to the scheme; on looking up a term, the user would find a class number, but possibly also some related terms. If we are asked for information on 'Thinners', we turn to the thesaurus and find

**Thinners** use  
**Solvents**

At Solvents we find

<b>Solvents</b>	XHG
UF	Thinners
RT	Dispersants
	...
	Solvent extraction
NT(A)	Paint thinners
	Turpentine

At HXG in the classification schedules we find

HX	<b>Materials by purpose</b>
HX2	Additives
HXG	Solvents

Looking at the display in the thesaurus and classification, we can find related terms as well as the terms we first think of, but with the classification schedule to help us by displaying the relationships in the broader context. The importance of thesaurofacet lay in the combination of two tools, but it was also welcomed in the UK as having a British bias. Though it has not been kept up to date it served an important function in emphasizing the relationship of thesaurus and classification.

### 7.31 BSI ROOT Thesaurus:

In 1966, the British Standards Institution began a service to exporters which drew on its large collection of overseas standards. In order to do so BSI devised a thesaurus based on 'Thesaurofacet'. In 1977 the need for an expanded and revised tool was seen to provide a standard vocabulary for ISONET, the international network of standards organizations. In 1981 BSI published the first edition of its ROOT Thesaurus, which was adopted by ISONET. The second edition was published in 1985 and the third in 1988.

Like thesaurofacet, the scheme has as its main basis a classification scheme, with a complementary alphabetical scheme, computer produced from the main schedules. The notation reflecting the structure is independent of language and in consequence the scheme lends itself to the production of editions in languages other than English. The scheme comes in two substantial volumes. The first of these is the subject display which lists over 12000 preferred terms and around 5500 non-preferred entry terms in a classified arrangement. Facet analysis has been generally used to show the structure of each subject. Because the scheme is intended to be used internationally, the usual BT-NT-RT codes have been replaced by symbols which are language independent and reasonably self-explanatory.

Symbol	Meaning
<	BT
>	NT
-	RT
*<	BT from another hierarchy
*>	NT from another hierarchy
*-	RT from another hierarchy
=	UF
→	USE
+	Used between two terms
**	Synthesized term
=**	precedes a noun-preferred term
[...]	SN
(By ...)	facet indicator

The notation used consists of capital letters in blocks of up to three separated by a period. A spread of notation is indicated by the slash (/) as in UDC. The notation is not intended to be rigidly hierarchical. The display begins with the subject contents list. This is an overall outline of the scheme. The classes are divided into two sections: the core which is the main purpose of the list and ancillary subjects – such as social sciences and humanities, which can be expanded in the future as the need arises. Ancillary subjects are marked + in the list (not in the thesaurus).

	Pages
+A General section	20
B Measurement, testing and instruments	36
+C/E Science	127
...	
+Z Social sciences and humanities	10

An example extracted from the schedule:

<b>A</b> <b>AP/AW</b>  AQ/AR  AQC AQE	<b>General section</b> <b>Common terms</b> [Prohibited term. Use a more specific term] Time  *->Exposure time LPU *->Operating time MBC.DP *-Time measurement BI Dates (Calendar) Seasons  =Autumn =Spring(season) =Summer =Winter	<b>Major heading</b> <b>Next step of division</b> Instruction (Prohibition) Next step of division NT in another hierarchy NT in another hierarchy RT in another hierarchy Indented subdivision Indented subdivision UF UF Homograph qualified Alphabeticalorder, not Chronological
---	--	--

The system obviously represents a considerable effort to match a particular objective that of providing a standard thesaurus for organizations concerned with standards. It has been used by other bodies as a source to enable them to compile their own more specialized thesauri, and is regularly updated by BSI. It is a good example of the way that a thesaurus based on systematic arrangement can be the basis of multilingual use.

#### 7.4 CLASSARUS:

Research carried out at the Documentation Research Training Center, Bangalore has led to three distinct but interrelated contributions: a General Theory of Subject Indexing Language (SIL); an indexing system known as Postulate-based Permu-term Subject Indexing (POPSI) and classarus. Both classification systems and thesauri have their specific strengths and weaknesses. Combining both approaches elimination of the later can preserve the strengths. Classarus, which originate in this well-known way, are most effective if they are constructed and applied during computer-aided indexing. Classarus is characterized by the employment of simple but highly effective conceptual and technical devices and by the renunciation of attempts to generate the wording of index entries algorithmically. Classarus is a faceted hierarchic scheme of terms with vocabulary control features designed on the basis of the General theory of SIL. It is a system of terms having separate hierarchic schedules of the Elementary Categories: Discipline, Entity, Property, and Action, together with their respective Species/Types, Parts and Special Modifiers. Also there are separate schedules for the Common Modifiers: Form, Time, Environment and Place. Each of the terms in these hierarchic schedules is enriched with synonyms and quasi synonyms. The hierarchic schedules constituting the systematic part are supplemented by an alphabetical index of chain entries. A classarus is used in the formulation of subject headings in general and in particular, subject headings according to the Postulate-based Permu-term Subject Indexing Language. For the construction of a classarus the POPSI language itself provides guidelines.

#### 7.5. Check your progress

##### 1.Cross- Reference is.....

**Ans:-**Cross-references are made for the purposes of guiding the users from their entry vocabulary to valid headings and linking related headings.

2. Sears' List of Subject Headings which was based on LC list (True/False)

Ans:- True

3. Who is the publishers of Sear's List of Subject Headings

(a) Dr. S.R.Ranganathan (b) J.D.Brown (c) **H.W.Wilson** (d)

C.A.Cutter

4. MESH is an example of

(i) classification scheme (ii) **thesaurus** (iii) abstracting journal (iv)

classaurus.

5. What is Thesaurus?

(a) A collection of selected terminology (b) Synonym terms (c)

List of words (d) **All of the above**

### 7.6 Summary

In this Unit you have been introduced to two major subject headings lists like Library of Congress Subject Headings, and Sears List of Subject headings. The concept of thesaurus, steps involved in its construction and two major thesauri in the field of science and technology, Medical Subject Headings (MeSH) and INSPEC Thesaurus have also been discussed in this Unit. The concept of thesaurusfacet including BSI ROOT thesaurus have been described in this Unit. Also, the concept of classaurus has been explained to you in brief.

### 7.7. Questions for self study

1. Highlight the functions and principles of LCSH
2. Mention the different phrase headings used in LCSH
3. How cross-references are given in LCSH
4. Explain the scope of the Sears List of Subject Headings
5. Bring out the major features of 18<sup>th</sup> edition of Sears List
6. Write the merits and demerits of Sears List of Subject Headings.
7. Define 'Thesaurus'
8. What is in a thesaurus?
9. Mention the general steps to be taken in thesaurus construction
10. Write briefly on collecting terms
11. When to use NT,BT and RT relationships?

## 7.8 REFERENCES:

1. Aitchison, Jean and Alan Gilchrist. *Thesaurus Construction: A Practical Manual*. 2nd ed. London: Aslib, 1987.
2. Chakraborty, A R and Chkraborthy, Bhubaneswar. *Indexing: Principles, Process and Product*. Calcutta: World Press, 1984
3. Chan, L. M. *Library of congress subject headings: Principles and application. Fourth Edition*. Westport, Conn: Libraries Unlimited. (Library and Information Science Text Series), 2005
4. Foskett, A.C. *The Subject Approach to Information*. 5<sup>th</sup> ed. London: LA Publishing, 1996.
5. Fundamental principles of Library of Congress Subject Headings: <http://www.itsmarc.com/crs/shed0161.htm> [Accessed on 28.2.2007]
6. Ghosh, S K and Satpathi, J N. *Subject Indexing Systems: Concepts, Methods and Techniques*. Calcutta: IASLIC, 1998
7. Hutchins, W. J. *Languages of indexing and classification. A linguistic study of structures and functions*. London: Peter Peregrinus, 1975
8. Lancaster, F. W. *Vocabulary Control for Information Retrieval*. 2nd ed. Info Resources Press, 1986.
9. Lancaster, F. W. *Indexing and abstracting in theory and practice*. London: Facet Publishing, 2003
10. <http://www.marisol.com/>
11. Langridge, D.W. *Subject Analysis: Principles and Procedures*. London: Bowker-Saur, 1989.
12. Miller, U. *Thesaurus and New Information Environment*. IN: *Encyclopedia of Library and Information Science*. New York: Marcel Dekker, 2003 pp. 2811-2819.
13. Miller, J. & Sears, M. E. (Eds.). *Sears list of subject headings*. 18th edition. New York: H. W. Wilson C, 2004
14. Prasher, R.G. *Index and Indexing Systems*. Ludhiana: Medallion Press, 1989
15. Riaz, Mohammad. *Advanced Indexing and Abstracting Practices*. New Delhi: 1989
16. Vickery, B.C. *Faceted Classification Schemes*. Rutgers Series on Systems for the Intellectual Organization of Information. Vol. 5. Ed. Susan Artandi. New Brunswick, New Jersey: Graduate School of Library Service, Rutgers, the State University, 1966.
17. *Wikipedia. The free encyclopedia*. (2006). *Library of Congress Subject Headings*. [http://en.wikipedia.org/wiki/Library\\_of\\_Congress\\_Subject\\_Headings](http://en.wikipedia.org/wiki/Library_of_Congress_Subject_Headings) [Accessed on 25.2.2007]

---

**MLISc – 5**  
**Information Systems: Architecture and Retrieval**

---

---

**Block – 2**  
**Practical and Strategic Issues in Digitization of Library Collections**

---

---

**Unit – 8**  
**Introduction to digital libraries**

---

**8.0 Objectives:**

**8.1 Definitions:**

**8.2 Current scenario of library systems**

**8.3 Evolution of Digital Library**

**8.3.1 World Brain**

**8.3.2 Memex**

**8.3.3 Thinking Center**

**8.3.4 Library Surrogates**

**8.3.5 Heritage Books**

**8.3.6 Local Culture and Language Preservation**

**8.3.6 Advances in ICAT**

**8.3.7 Networks**

**8.3.8 Metadata Standards and Formats**

**8.3.9 Enabling Technology**

**8.4 Internet:**

**8.4.1 Popularity of Internet**

**8.5 Presentation Technology**

**8.5.1 Unicode:**

**8.6 The benefits of digitization**

**8.7 Components of Digital Library**

**8.7.1 Data**

**8.7.2 Self Check Exercise:**

**8.7.2.1 Information store i.e., an information base. This has two components (digital infrastructure)**

**8.7.2.2 Content Creation and capturing the content**

**8.7.2.3 Search and Access Mechanism**

**8.8 Practical and strategic issues in the digitization of library collections**

**8.9 An Overview of Digitization Projects:**

**8.9.1 Digitizing Gutenberg**

**8.9.2 The Digital Shikshapatri**

**8.9.3 The International Dunhuang Project**

**8.9.4 The New York Public Library**

**8.9.5 New Initiative**

**8.10 Check your progress**

**8.11. Summary**

**8.12. Glossary**

**8.13. Questions for self study**

**8.14. References**

**8.0 Objectives:**

- ❖ To know how the concept of digital library is defined is it analogous to electronic, Highbred and Virtual library.
- ❖ To know the landmark events that let to the digital library initiatives.
- ❖ To understand the maps components of digital library.
- ❖ Content Creation and capturing the content
- ❖ An Overview of Digitization Projects

**8.1 Definitions:**

In the literature the terms digital library and electronic library are used interchangeably, though the latter is more popular in the UK Borgman (1999) has analysed the various definitions and connotations of digital libraries that have been proposed by researchers throughout the world. She argues that the research community's definition of digital libraries has evolved from a narrower view emphasizing technologies to a border view encompassing the social, behavioural and economic contexts in which digital libraries are used. In her opinion, the views of the library community are reflected

in the definition given by the Digital Library Federation (DLF) as reported by Waters (1998): ‘Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities’.

In addition, a new concept, the hybrid library, has also emerged (Pinfield et al., 1998; Rusbridge, 1998; Chowdhury and Chowdhury, 1999; Oppenheim and Smithson, 1999). Rusbridge (1998) suggests that a hybrid library brings a range of technologies from different sources together, and it integrates systems and services in both the electronic and print environments. He further argues that ‘the name hybrid library is intended to reflect the transitional state of the library, which today can neither be fully print nor fully digital,. Pinfield et al., (1998) suggest that the hybrid library is the continuum between the conventional and digital library.

The type of information that a digital library handles ranges from text, numerical data, figures, photographs, maps, slides, to music, video and films. Digital libraries differ from traditional libraries in particular ways, and many publications have discussed these characteristics (Gladney et al., 1994; Vicki and Winograd, 1995; Chowdhury and Chowdhury, 1999). Some of these important characteristics are:

- information resources can vary from simple text to multimedia available at one or several locations; they may be available on different platforms, and may have been created/ or organized differently
- information may come from various sources- from electronic journal producers or vendors to databases; from local digital libraries to remote digital libraries; and so on
- digital materials form part of a larger collection that comprises print materials
- information may be coupled with complex metadata structures
- users can be located anywhere and their nature, information needs, etc., may vary significantly

- there is no human intermediary and no physical collection, at least at the point of interaction
- a range of services, such as searching filtering and downloading, as well as current awareness and selective dissemination of information services, may be provided
- there are many complex issues of information retrieval, access management control of intellectual property rights, security, authentication, etc.
- in many cases information is not owned; only a right to access is provided
- there may be several versions of the same information.

### **Digital libraries**

The digital library, the electronic library (generally taken to synonymous with the digital library), the virtual library, the hybrid library, the library without walls are all concepts that librarians seem to be dealing with all the time. What do they mean? Do they mean the same to everyone who uses the terms? Do they mean the same thing? Do we all mean the same thing when we talk about a library? From its original etymological meaning of a ‘collection of books’, a library can mean a collection of almost anything in modern parlance: software routines, for instance.

### **Arms (2000, 2), defines a digital library as:**

A managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network. A crucial part of this definition is that the information is managed.

Borgman (2000) devotes a whole chapter to trying to define a digital library (called ‘Is it digital or is it a library?’) and concludes that the term has multiple meanings, and, for her, these cluster around two themes:

From a research perspective, digital libraries are content collected and organized on behalf of user communities. From a library-practice perspective, digital libraries are institutions or organizations that provide information services in digital forms.

The digital is one more (albeit very different) format that librarians have to deal with in a multi-format environment, where for many years non-documentary mixed-media objects, both analogue and digital, have been growing in importance, and where technology for access had to be provided.

### **According to Rusbridge (1998)**

Designed to bring a range of technologies from different sources together in the context of a working library, and also to begin to explore integrated systems and services in both the electronic and print environments. They exist on the continuum between the conventional and digital library, where electronic and paper-based information sources are used alongside each other' (Pinfield et al., 1998, 3). The concepts of virtual libraries or libraries without walls reflect the integrative possibilities inherent in the digital: if significant library collections are digital, then the confines of space no longer define boundaries upon information. Virtual collections from many different sources can be assembled and accessed from anywhere, without the user even knowing where the sources reside, and personal virtual collections can be built to serve many purposes. Hence a virtual library could potentially be enormous, linking huge collections from all round the world together, or it could be very small being the personal digital collection of one individual.

Digital libraries are like any so-called revolutionary change, a development of a whole range of underlying theories and technologies that have come together to create a paradigm shift. The speed of recent developments has taken some librarians by surprise, especially the exponential growth of the amount of digital data available, but they are understandable if we look at some of the precursors that have led to the current trends: new developments rarely spring fully formed from the ether.

There are many different kinds of digital libraries: Creating, delivering and preserving digital objects that derive from many different formats of underlying data, and it is very difficult to formulate a definition that encapsulates all these.

Based on the definition above we can characterize digital libraries as follows:

- 1 A digital library is a managed collection of digital objects.
- 2 The digital objects are created or collected according to principles of collection development.
- 3 The digital objects are made available in a cohesive manner, supported by services necessary to allow users to retrieve and exploit the resources just as they would any other library materials.
- 4 The digital objects are treated as longer as long-term stable resources and appropriate processes are applied to them to ensure their quality and survivability.

### **Virtual Library**

The concept of virtual libraries or libraries without walls reflect the integrative possibilities inherent in the digital if significant library collections are digital, then the confines of space no longer define boundaries upon information. Virtual collections form many different sources can be assembled and accessed from anywhere, without the user even knowing where the sources reside, and personal virtual collections can be built to serve many purposes. Hence a virtual library could potentially be enormous; linking huge collections from all around the world together or it could be very small, being the personal digital collection of one individual.

Virtual library therefore provides a portal to information that is available electronically elsewhere. The source information available electronically can be readily displayed with a web browser. A good example of virtual library is INFOMINE, a co-operative project of university of California and California state university (among others). Run by librarians, it covers most major academic display through access to important data bases, e-journals, e-text and other digital corrections.

### **8.2 Current scenario of library systems**

Barker has formulated four basic possibilities of library systems. They are

- 1). Poly media libraries;
- 2). Electronic libraries;
- 3). Digital libraries and
- 4). Virtual libraries.

Poly media denotes the use of several different, independent media for the storage of information and knowledge: paper, microfilm, compact disc, and so on. Thus, poly media libraries are institutions that store their material on a wide range of independent media types. They are similar to conventional libraries and their organizational and management process are basically manual in nature.

In an Electronic library the core processes are basically electronic in nature through the widespread incorporation and use of computers and through various facilities such as on-line indexes, full text searching and retrieval facilities automated record keeping and computer based decision making. In addition, within an electronic library system there will be a conscious movement toward the more wide spread use of electronic media (both digital and analogue) for the storage, retrieval and delivery of information.

In digital libraries the information contain only, in a digital electronic format. Of course, the information itself may reside on different storage media such as electronic memory or magnetic and optical disk, but users will not perceive any difference between them. Obviously, in order to access digital information it is necessary to use either special purpose, multimedia 'reader' stations or some form of computer system. One advantage of having information in digital form is that it can be shared simultaneously and easily at relatively low cost therefore, a digital library could generate an unlimited number of copies 'at the touch of a button'.

Virtual library systems depend for their existence of a technology known as virtual reality. This has been described as 'ultimate multimedia experience and depend upon the ability of a computer and its associated interaction peripherals to create highly realistic simulations and surrogations in which users can become totally immersed.

The virtual libraries provide facilities to support this type of experience such libraries provide facilities to browse around a library system without actually having to physically go to it.

Although some virtual library systems are available in the form of packaged CD-ROM products, the most advanced systems that are supported by advanced telecommunication equipments to facilitate remote access and sharing such libraries can be accessed using two dimensioned (2D) interfaces based upon conventional computer interfaces or 3D interfaces involving the use of head-mounted displays and peripherals such as a 'data glove'. These facilitate the creation of a 'total immersion' environment and interaction based upon pointing operations and gestural communication. Using equipment of this sort it is possible to enter a virtual library, browse around its rooms and shelves, use an index or catalogue, select a book (by pointing to it or touching it) open it and then read it. Of course, the only place where the book really exists is in the computer and in the minds of its readers.

### **8.3 Evolution of Digital Library**

#### **8.3.1 World Brain**

The concept of Digital Library is not new. Over 60 years ago, science fiction writer H.G. Wells was promoting the concept of a "World Brain" based on a permanent world encyclopedia which "would be the mental background of every intelligent (person) in the world. It would be alive and growing and changing continuously under revision, extension, and replacement from the original thinkers in the world everywhere" (Wells 1938).

The idea of H.G. Wells was an attempt to develop a global depository of knowledge. This also has propelled many technologies to realize his dream of universal access to universe of knowledge.

#### **8.3.2 Memex**

Eight years later ‘Vannevar Bush’, the scientific advisor in the US War Force argued the need for

- a. Future device for individual which is a sort of mechanized private library, a device in which an individual stores all his Books, records, and communications, so that it may be consulted with exceeding speed and publicity”. The automated library as envisaged by Bush is called the Memex (Bush, 1947).

J.C.R. Licklider, of the U.S. Department of Defence, envisioned that human brains and computing machines would be tightly coupled together and supported by a “network of ‘Thinking centre’ that will incorporate the functions of present day libraries together with anticipated advances in information storage and retrieval” (Licklider, 1960).

### **8.3.3 Thinking Center**

Wells, Bush, Licklider and other visionary thinkers were advocating something very close to what we might now call a virtual library. Virtual library is something in existence in essence or effect though not in actual fact, form or name. A virtual library is a library for all practical purposes, but library without walls or books.

### **8.3.4 Library Surrogates**

A library catalogue is a complete model that represents, a summary of, if not a surrogate for, library contents. Today we call this “metadata”.

Other essential aids to information seeking in libraries are published bibliographies and indexes. Like catalogues these are virtual representations – metadata – and they provide the traditional means of gaining access to journal articles, govt. documents, microfiche, microfilm and special collections. The information in library catalogs and bibliographies can be divided into two kinds; the first having reference to the contents of books, the second treating their external character and the history of particular copies.

### **8.3.5 Heritage Books**

Beautiful books are costly for their splendid illustrations, decorated illuminated letters, and are being printed on uncommon paper, or uncommon materials. They have unusual bindings and are rare and having historic significance. In India, there are ancient books, some are almost 2000 years' old written on palm leaves, bound with string threaded through holes in the leaves. Whimsical bindings abound for example in fox skin, in human skin, deerskin edged with gold tooling and the embossment. These are to be preserved digitally. Digitalization also help disseminating humanitarian information and save costly and rare documents from disasters, such as earthquakes or hurricanes, terrorist attacks or nuclear accidents and Documents can be digitized and in the information continued therein can be organized and distributed even in the absence of an effective network infrastructure.

### **8.3.6 Local Culture and Language Preservation**

Digitalization also helps to preserve our thought, communication and culture identity; individual medical knowledge based on local plants or long acquired knowledge of the cultivation and protection of local species. Can all be preserved digitally? This will allow developing countries to participate actively in Information. Society, rather than observing it from outside. It will stimulate the creation of new industry. It will help ensure that intellectual property remain where it belongs, in the hands of those who produce it.

### **8.3.6 Advances in ICAT**

Computer technology, communication technology and the presentation technology have propelled the development of digital libraries. Personal computers are found almost in every house in the developed countries. In 1998 World Bank Survey of developing countries found 3 to 30 PCs per 100 people, depending on the poverty level with growth prediction at 20% per year. It is estimated that at the turn of the millennium there will be 50 million PCs in developing countries serving a population of 4 billion.

### **8.3.7 Networks**

Network access is widely around the globe. In 1998 more than a quarter of the population was surfing in the Internet. With the global Satellite communication networks the developing countries would eventually have greater network access. Digitization activity will help create the collections.

### **8.3.8 Metadata Standards and Formats**

The data presentation may be traced to 1930's, 40s, 50s, and 60's from the time of Memex or even to early information retrieval system to the genesis of Hypertext system (1960's and 70's) SGML and metadata formats (1990's). The major technologies are the building blocks of the digital library: communication technologies, digital computers and digital storage devices. The power of the computer to store organizes and retrieve is immensely stupendous and progressing.

### **8.3.9 Enabling Technology**

Before the days of networking the information storage in individual computers were available only to those who had direct access to those with a desktop. But what transformed this scenario to a situation wherein the information stored on the desktop is accessible to any one sitting any where in any time, through the networking technology and the developments in the software technologies. When we talk about technology in the context of digital library we talk about both hardware and software side of the digital technology. So far as network is concerned it is not just the Hardware that enables access, but it is the software part of the technology. Mention may be made of IP / TCP (internet Protocol/ Transfer Control Protocol). That is transmission control protocol. This protocol has enormously contributed to the provision of global access to information.

## **8.4 Internet:**

The Internet has given us the means for creating digital libraries and for making them accessible. During 1990's many experiments were conducted and resources on a particular subject (collections) in an individual library will bring together on web pages at each cooperating institutions. But "digital library" quickly came to mean collections in which a site provides digitized information resources with an architecture and a service for the retrieval of such resources Christine Borgman (2000) emphasized that digital

libraries are for communities of users and that they are really are extensions of the physical places where resources are selected, collected organized, preserved and accessed, including libraries, museums, archives and schools. Organization of digital libraries is being accomplished with such tools as metadata, XML/RDF schemas, ontologies and taxonomies. You are given sufficient briefings about them in the units that follow in this Block.

#### **8.4.1 Popularity of Internet**

Internet began in 1969, when the ARPANet was first and experimented in Military communications even during a nuclear attack. In United States Paul Barren of the band corporation and in U.K, Donald Devis of National Physical Laboratory independently developed the same principle of communication, which Davies termed as packet switching the Department of Defence in the United States which founded the Advance Research Projects Agencies worked on the experimental. This method of communication under a project ARPANET packet switching is a form of digital communications where network traffic is broken-up into small information packets, each one put, units own digital envelop and each separately addressed and sent separately (routed) through the network. You have learned already under course – 6 Fundamentals IT about Internet during your BLISc programme. You are directed to study more on the infrastructure for network communications. There are communication internet protocols. These protocols are used for sending data communication on data networks, particularly over long distances. Internet serves as the courier and provides the content delivery mechanism. It a commonly known as TCP/ IP, transmission control protocol and Internet Protocol. There are seven important factors which made internet popular. These are

1. ASCII: all communications on the internet are built on the ASCII standard
2. Packet Switching
3. Protocols: Protocols are the rules of communicating, so the Internet Protocol (IP) describes what is in the IP packet header and how that information is organized so that all nodes in the network have the same understanding of what that information means.

4. Open Architecture: the internet networking protocols are in the public domain so that anyone can implement them.
5. Client – server: the internet is designed on the client server model which is a design for software that minimizes network traffic and allows that internet to run on many kinds of machines.
6. Cross Platform: The Internet allows computers of all most all types of communication each other TCP / IP is open source software and anyone can develop and use and run it on any platform.
7. Bottom up development: As result of these factors, a lot of people got involved in the development of internet. Rather then being developed by one company in the proper interest in the development and control of a suite of protocols.
8. The internet has enabled global connectivity of computer networks and the development of various tools and techniques for networked information provision and access.

### **The Web:**

The Web the most variable information service provides a very convenient means for publishing and accessing multi-media hypertext linked documents stored in the computers spread across the word. “Web has brought to the desk-top, not only metadata sources like bibliographic databases and table of content, but also full text of journals, preprints, technical reports, patents, course work etc. On the web, information is just a ‘click away’. Perhaps this was the dream Vannevar Bush had when no proposed ‘Memex’ over five decades ago and more recently ‘Xamades’ by Ted Nelson.”. The Web provides the tools and techniques for content publishing hosting and access. Digital libraries are much more than this.

### **8.5 Presentation Technology**

Today internet has become part of our life cycle. This protocol helps the flow of information, flow of signals through the hierarchy. Internet is nothing but network of networks. TCP / IP help an unimpeded information flow globally. The other technology which made digital library possible is known as presentation technology. Which hold streams of ideas in terms of text, image, video and sounds. These streams of ideas are

transformed into the computer language known as bits and bytes. Presentation technology make the streams of bit into text, picture, sound etc, and provide universal access to information. ASCII, (American Standard Code of Information Interchange) is a standard 7 bit code for representing the Roman alphabet, numerical and special symbols.

Textual objects are generally formatted in the international ASCII standard for representing character formats or one of the ASCII extensions for encoding diacritical characters in romance languages other than English.

A bit is an electronic impulse that can be represented by two states, 'on' or 'off' also terms as '1' or '0' A 'Bite' consists of 8 bits and 1 bite represents 1 alpha numeric character. A ten-letter word for example would be ten bytes. Bits and bytes are linked together in chains of millions of electronic impulses; this is known as the 'bit stream' A 'Kilobyte' is 1024 byte, and a 'megabyte' 1024 kilobyte. Digital images are represented by 'pixels' or picture elements - dots on the computer screen or printed on paper. Pixels can carry a range of values, but at the simplest level, one pixel equals one bit, and is represented in binary form as 'black' (off) or 'white' (on). Almost any kind of information can be represented in these seemingly simple structures, as patterns of the most intricate complexity can be built up. Most primary sources held in libraries and other cultural institutions are capable of digital representation (anything that can be photographed can be digitized, though these are issues with the faithful representation of 3D objects) When digital they are susceptible to manipulation, interrogation, transmission and cross linking in ways that are beyond the capacity of analog media. Alphanumeric symbols are the easiest objects to represent in digital form, and digital text has been around for as long as these have been stored – program computers. These symbols are also most compact to store, an important factor at a time when capacity was limited and expensive and were good at a time when capacity was limited and expensive. Early computers were good at processing symbols rapidly but input, output and display of data were difficult.

As the Internet exploded into the World Wide Web and burst into all countries and all corners of our lives, the situation became untenable. The world needed a new way of representing text.

### **8.5.1 Unicode:**

In 1988 Apple and Xerox began work on Unicode, a successor to ASCII that aimed to represent all the characters used in all the world's languages. As word spread, a consortium of interested parties was formed in 1991. The result was a new standard, ISO – 10646, ratified by the International Standards organization in 1993. Unicode continues to evolve; the main goal of representing the scripts of languages in use around the world has been achieved. There is a steady stream of addition, clarifications, and amendments which eventually lead to new published versions of the standard. Recent programming languages – notably Java have built in Unicode Support. All principal operating systems support Unicode, and application programmes, including web browsers; have passed on the benefits to the end user. Unicode is the default encoding for HTML and XML. People of the world, rejoice. Unicode is universal. But it also satisfies a stronger requirement in the resulting Unicode file can be mapped back to the original character set without – any loss of information. This requirement is called round trip compatibility with existing coding schemes, and it is central to Unicode's design.

### **Search features of selected digital libraries**

There are different categories of digital libraries, and their search and retrieval characteristics depend on their nature, information content, target users, and so on. The features of digital libraries:

- digital libraries that deal with specific subjects and types of materials, NCSTRL and NDLTD
- a digital library that deals with non-textual information, the Alexandria digital library
- a digital library that deals with different types of materials, NZDL
- Hybrid library that handle digital and well as traditional library materials, CDL and Headline.

## 8.6 The benefits of digitization

The digitization of resources opens up new modes of use, enables a much wider potential audience and gives a renewed means of viewing our cultural heritage. These advantages may outweigh the difficulties and disadvantages; provided the project is well thought out. Institutions large and small are therefore embarking upon programmes of digital conversion for a whole range of reasons. The advantages of digital surrogates include:

- immediate access to high-demand and frequently used items
- easier access to individual components within items (e.g. articles within journals)
- access to materials held remotely
- rapid the ability to reinstate out-of-print materials
- the potential to display materials that are in inaccessible formats, for instance, large volumes or maps
- ‘virtual reunification’ – allowing dispersed collections to be brought together
- the ability to enhance digital images in terms of size, sharpness, colour contrast, noise reduction, etc.
- the potential to conserve fragile/precious originals while presenting surrogates in more accessible forms
- the potential for integration into teaching materials
- enhanced searchability, including full text
- Integration of different media (images, sounds, video, etc.)
- The ability to satisfy requests for surrogates (photocopies, photographic prints, slides, etc.)
- reducing the burden or cost of delivery
- The potential for presenting a critical mass of materials.

## **8.7 Components of Digital Library**

The main component of digital library may be as follows:

### **8.7.1 Data**

This consists of books and a journal stored in a digital form on a computer disk store. The way of storing this information; one way is to photograph a page and scan the image with a scanner. This form of storage is called a bit mapped form. It is a practical way of storing old books and journals. The image of a page may be retrieved and displayed on the video screen of the computer.

#### **Audio Data**

Audio data (audio files) are digitized; compressed using commonly accepted standard compression algorithms and stored. Musical may also be coded and stored along with audio data.

#### **Video Data**

This requires enormous storage space due to the need for repeating frames at least 30 times per second. Thus the data are compressed in such a way that when decompressed the original data are recovered.

#### **Linking**

The information collection of the digital library will normally not be stored in one computer rather it will be distributed in many computers known as servers. All those servers will be linked by high-speed communication links. The fact that the information in the digital library is distributed need not be known to a user as it is not relevant from his point of view. A user gets easy access to the information based on his request regardless of its geographical location.

#### **User**

The user can access the library from anywhere using a terminal or a computer connected to the network to which the information servers of the library are connected.

#### **Indexing**

Indexing and inter linking multimedia data are extremely important for ease of retrieval. Keywords in textual documents are selected and linked to related words with logical links by appropriate software's through hypertext links.

### **Photograph**

Color and monochrome photos are stored in bit-mapped form using compression algorithms to reduce storage space.

### **Numeric Data**

These consists of tables of various types such as a physical property of data of various materials, data form experiments and astronomical tables etc. such numeric data stored digitally may be used by curve setting programmes, spreadsheet programmes etc.

### **Formats of material for digitization**

Digitization is a process of conversion of any physical or analogue item into a digital representation or facsimile. The physical items that may be candidates for digitization may include:

- individual documents
- bound volumes, both print and manuscript
- photographs, -both prints and transparencies
- microfilm and microfiche
- video and audio
- maps, drawings and other large-format paper items
- art works
- textiles
- Physical three-dimensional (3-D) objects.

The digitization processes are numerous, and include:

- image scanning
- microfilming and then scanning the microfilm
- photography followed by scanning of the photographic surrogates
- re-recording video and audio on to digital media

- rekeying of textual content
- OCR of scanned textual content
- tagging text and other digital content to create a marked-up digital resource
- Digital photography – especially for 3-D objects or large-format items such as art works.
- Sathyanarayana identified the following five components of the digital library

### **8.7.2 Information store i.e an information base. This has two components (digital infrastructure)**

- a) Metadata and
- b) Digital library objects (primary documents) Metadata (catalogue or surrogate) facilitates identification and location of digital objects using a variety of search techniques and provides access to the digital objects. Digital objects are the primary documents required by the users. These may be stored in wide range of digital formats: text, image, audio and video content. These may be stored in the same server that host the Metadata, or may be distributed over different servers. However, most digital libraries today are internet based. Several standards and formats for efficient storing and delivery digital information over the internet. AACR, MARC, CCF and the latest Dublin Core are some the standards developed by the profession to describe the data elements. Format for handling primary elements include ASCII, HTML and SGML for text, PDF for scanned documents, GIF and JPEG for images, Real Audio and MP3 for audio, MPEG for Video, XML (eXtensible Markup Language).

#### **8.7.2.1 Content Creation and capturing the content**

Digital content may have to be created newly or scanned or converted from paper or digitally captured from artifacts or extracted from archival documents. The technology tools for content creation and capture include scanners, digital camorras audio and video cards from analogue to digital audio and video. OCR packages help extracting text from scanned documents. We also need software packages for converting from one digital format to another. A key requirement in content creations the preparation of Metadata

providing description and induce information for each digital object, facilitating easy search and retrieval.

### **8.7.2.2 Search and Access Mechanism**

This includes the user interface for searching and display interface for showing search results. Typically most digital library support both browsing and searching. 'Browsing is menu driven for selecting the item of search (by subject, by author or by title, hierarchical). Searching is by 'querying' wherein the user poses an explicit search query. Digital library support searching with varying degree of capabilities, including simple word based search, Boolean search, phrase searches, field based searches etc. apart from Metadata searching, full text searching is also often supported. A few digital libraries also support searching digital libraries may also support federated searches.

### **Activities involved in digitization ?**

Any digitization project is likely to involve some or all of the following activities:

- assessment and selection of originals
- grant applications and fundraising
- feasibility testing, costing and piloting
- copyright clearance and rights management
- preparation of materials
- benchmarking
- digital capture
- quality assessment
- metadata design and creation
- delivery
- workflow processes
- project management
- long-term preservation.

## **8.8 Practical and strategic issues in the digitization of library collections**

Digital collection development is part of a border perspective on collection development, and generally needs to be assessed using the same criteria. However, there is a difference between reviewing collections already held by the institution with digitization in mind, and choosing to acquire digital materials from elsewhere. We deal with building digital collections in more detail in Chapter 3 ‘Developing collections in the digital world’. In-house holdings have some value to the community served by the library, and thus theoretically they might all be candidates for library and the community it serves that will dictate certain priorities for digital capture and delivery. As Hazen, Horrell and Merrill-Oldham (1998) warn:

The judgments we must make in defining digital projects involve the following factors: the intellectual and physical nature of the source materials; the number and location of current and potential users; the current and potential nature of use; the format and nature of proposed digital product and how it will be described, delivered, and projections of costs in relation to benefits. The typing of material available for potential digitization in any library (however small) is likely to be greater than the resources available and so careful assessment of costs and benefits need to be made before embarking on projects.

## **8.9 An Overview of Digitization Projects:**

### **8.9.1 Digitizing Gutenberg**

Johannes Gutenberg and the Bibles he printed in the mid-15<sup>th</sup> century have iconic significance in the worlds of both written and digital culture.

Given the pivotal cultural significance of Gutenberg's work, it is appropriate that his surviving Bibles should be viewed as candidates for digitization. But there are many other reasons for producing digital facsimiles of this work, not the least being its aesthetic qualities and the possibility of bringing these to a wider audience. In 1996, Keio University in Japan embarked on an ambitious programme to capture digital facsimiles of Gutenberg Bibles, including its own volume acquired that year. The HUMI Project (Humanities Media Interface, and also a Japanese word for books, writing or learning) plans to digitize ten copies in all, and has currently completed six of these: the Keio copy,

two copies at the Gutenberg Museum in Mainz, two copies at the British Library and one copy in Cambridge University Library.

### **8.9.2 The Digital Shikshapatri**

The Digital Shikshapatri is a joint project of the Indian Institute Library and the Refugee Studies Centre at Oxford University and the Oxford Centre for Vaishnava and Hindu Studies. It has the wide participation of Swaminarayan Hindus from around the world. The main aim of the project is to digitize the Indian Institute's Shikshapatri manuscript, one of the great treasures of British Hinduism, and many different supporting materials, and make these available on the internet and on CD-ROM for use by a wide range of individuals and institutions. The Shikshapatri text contains the essence of Hindu moral codes for everyday life, and is of huge importance for devotees of the Swaminarayan Hindu sect.

### **8.9.3 The International Dunhuang Project**

The International Dunhuang Project (<http://idp.bl.uk/>) is an excellent example of how international co-operation in digitization projects can create a resource to enhance the study of a wide range of valuable materials that are scattered all over the world.

The history of the Dunhuang Project has a whiff of Indiana Jones romance about it. The materials in question were discovered in a cave near Dunhuang in China in 1900 by a Daoist monk. The cave had been sealed up since the 11th century, and inside was a library of tens of thousands of Buddhist texts amassed by monks between the fifth and the 11th centuries, including manuscripts, scrolls and wooden tablets, written in many languages and scripts. Besides the written materials were textiles, wood-block prints and other artefacts. In 1907, the archaeologist Sir Aurel Stein was the first foreigner to visit the site, and soon thereafter other scholars from all over the world arrived, and departed with many of the treasures, which were thus dispersed all over the world. What was discovered over the next few years was a whole complex of caves filled with precious artefacts and with walls covered with exquisite paintings. Around 492 decorated caves, large and small, are extant today. Dunhuang, a small 2000-year-old town in northwestern China, was once an important caravan stop on the Silk Road linking Central Asia with

China, hence the richness and diversity of the collections. There is an excellent description of the caves and their discovery by Roderick Whitfield ([www.textile-art.com/dunl.html](http://www.textile-art.com/dunl.html)).

Stein brought thousands of manuscripts and other artefacts back to Britain, where they now reside in the British Library. Other materials are held in Beijing, Paris and St Petersburg, with considerable collections in Japan. Digital technologies offer the potential to reunite these materials virtually, and also to enhance them with images of the cave paintings, as well as editorial materials, annotations, transcriptions, translations, etc. The British Library now has a large searchable database of manuscript catalogue records and digital images online (<http://idp.bl.uk/IDP/idpdatabase.html>), with information about some 20,000 manuscripts and printed documents

#### **8.9.4 The New York Public Library**

The New York Public Library ([www.nypl.org](http://www.nypl.org)) has made available a number of its collections in digital form, concentrating in particular on materials deriving from local (New York City and State) and wider US sources.

#### **8.9.5 New Initiative**

In the USA, the Library of Congress has been leading a programme to create a National Digital Library, which began in the early 1990s. The Library worked with the National Science Foundation and a number of institutions funded by the Foundation, as well as publishers, museums and educational bodies in the USA and elsewhere. It now has some seven million digital items from more than 100 historical collections in its American Memory database and this continues to grow rapidly. The Library of Congress has been working with international partners to promote strategic digital library implementations worldwide, and has also been influential in the development and Promotion of methodologies, technologies and standards (<http://lcweb2.loc.gov/ammem/ammemhome.html>)

The Bibliotheque nationale de France began planning a major programme of digitization at the same time as plans were drawn up for the new library at the Francois Mitterrand/Tolbiac site at the end of the 1980s. The Bibliotheque now has around 30 million pages of documentary materials available in digital form (much of which relates to the 19th century)

The British Library also began conceiving of new digital possibilities while planning a new building, and digital technologies are integral to its operations now that the Library has moved to the new site. During the 1990s, the Library undertook a programme of technical innovation under its Initiatives for Access programme, intended to 'assist the library in intelligently appraising the best way to exploit the new opportunities that technology is increasingly offering' (Alexander and Prescott, 1998, 17).

The projects have included the digitization and advanced image enhancement of the Anglo-Saxon Beowulf manuscript, the digitization of Gandahran Buddhist scrolls - 2000-year-old artefacts which are the oldest South Asian manuscripts of any type - the digitization of sound archives, of patents and also of microfilm. Now that it is installed in its new St Pancras premises, the Library is embracing digital access to its collections and its new goal is that 'the collections of the British library and other great collections will be accessible on everyone's virtual bookshelf - at work, at school, at college, at home' ([www.bl.uk/](http://www.bl.uk/))

In Japan, the Electronic Library Service of the National Center for Science Information Systems (NACSIS-ELS) provides an integrated system of bibliographic databases and electronic document delivery of Japanese academic journals on the internet.

The New Zealand Digital Library (NZDL) is a research project at the University of Waikato whose aim is to develop the underlying technology for digital libraries and make it available publicly so that others can use it to create their own collections. A number of collections of New Zealand and Pacific materials are made available through

the waikato website, as well as collections in music, Arabic and women's studies, and a considerable amount of material in human rights.

It is not enough if you just have all the technologies, computer, telecommunications and presentation technology unless you have the organizational support in terms manpower, finance and other infrastructure. Several national and international agencies cooperative and consortia in USA took initiatives to develop digital libraries. Outstanding among them are Coalition for Networked Information (CNI), the Council on Library and Information Resources (CLIR), Digital Library Federation (DLF), the Institute for museum and Library Services (IMLS), National Science Foundation (NSF), Digital Libraries Initiatives (DLI), Online Computer Library Center (OCLC).

Several digital projects were launched in the last decades of the 20<sup>th</sup> century throughout the world. In India several university, university departments and public libraries took part in the 'Million Book Project'. Mention may be made of Andhra Public Library, Karnataka Directorate of Public Libraries, S V University Library (Thirupati) and Library and Information Science Department of University of Mysore, Sponsored a number of digitization projects Central Institute of South Indian Languages, Mysore.

## 8.9. Check your progress

1. Arrange the following according to their year of origin:

(i) UNICODE (ii) XML (iii) HTML (iv) SGML

Codes

(A) (iv), (iii), (i), (ii) (B) (iii), (iv), (ii), (i) (C) **(ii), (iv), (i), (iii)** (D) (i), (ii), (iii), (iv)

2. Mention Character encoding Standard.

Ans. Unicode.

3. Architecture of digital library is based on

(a) Green environment (b) Political environment (c) **distributed technology environment** (d) None of the above

4. SGML is an international standard that describes the relationship between a document's content and its structure. (True/ False)

Ans:- True

5. The first set of RDA vocabularies published on the

(i) OAI (ii) Metadata (iii) AACR2 (iv) **Open Metadata Registry**

## **8.10 Summary**

In addition, a new concept, the hybrid library, has also emerged (Pinfield et al., 1998; Rusbridge, 1998; Chowdhury and Chowdhury, 1999; Oppenheim and Smithson, 1999). Rusbridge (1998) suggests that a hybrid library brings a range of technologies from different sources together, and it integrates systems and services in both the electronic and print environments. He further argues that ‘the name hybrid library is intended to reflect the transitional state of the library, which today can neither be fully print nor fully digital,. Pinfield et al., (1998) suggest that the hybrid library is the continuum between the conventional and digital library.

## **8.11 Glossary**

### **Audio Data**

Audio data (audio files) are digitized; compressed using commonly accepted standard compression algorithms and stored

### **Video Data**

This requires enormous storage space due to the need for repeating frames at least 30 times per second.

### **Linking**

The information collection of the digital library will normally not to be stored in one computer rather it will be distributed in many computers known as servers

### **Numeric Data**

These consists of tables of various types such as a physical property of data of various materials, data form experiments and astronomical tables etc.

### **Photograph**

Color and monochrome photos are stored in bit–mapped form using compression algorithms to reduce storage space.

## 8.12 Questions for self study

1. Write a note on thesaurus
2. Bring out the features of BSI ROOT Thesaurus
3. What is a classarius?

## 8.13. References

1. Brophy, Peter (2001), The library in the twenty-first century: new services for the information age; Library association publishing, London.
2. Marilyn deegan. Simon tanner (2002), Digital futures, strategies for the information age; Library association publishing, London.
3. Borgman, C.L. (2000), From Gutenberg to the global information infrastructure: access to information in the networked world, MIT Press.
4. Chowdhury, G G and Chowdhury S (1999), Digital Library Research: major issues and trends, Journal of Documentation, 55 (4), 409-48.
5. Crawford, W (1999), Being analog: creating tomorrow's libraries, American Library Association.
6. Rowlands, I and Bawden, D (1999), Digital libraries: a conceptual framework, Libri, 49 (4), 192-202.
7. Rusbridge, C (1998), Towards hybrid library, D-Lib Magazine, (July-August) available at.
8. Sathyanarayana, N V (2005), Components of Digital Library. In Turning Pages: Reflections in Infotimes. Ed by Srinivas Bhogle. Pages 69-71. Bangalore. Informatics (India) Ltd.  
<http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlibjuly1998/rusbridge/07rusbridge.html>



**KARNATAKA STATE OPEN UNIVERSITY**  
**MUKTHAGANGOTRI, MYSORE –570 006**

Master of Library and Information Science  
**M.Lib.I.Sc - 5**

**Information Systems:  
Architecture and  
Retrieval**

**BLOCK - 3**

**BLOCK**

**3**

---

**Practical and Strategic Issues in Digitization of Library Collections**

---

---

**Unit – 9**

**Architectural Agents and Tools for Digital Libraries**

---

**Unit – 10**

**Multilingual and Multi script issues**

---

**Unit - 11**

**Human Computer Interfaces**

---

**Unit – 12**

**Resource Discovery**

---

## INSTRUCTIONAL DESIGN AND EDITORIAL COMMITTEE

### COURSE DESIGN

**Prof. D. Shivalingaiyah**

**Chairman**

Vice Chancellor  
Karnataka State Open University  
Mukthagangotri, Mysuru-570006

**Prof. M. Mahadevi**

**Convener**

Dean (Academic)  
Karnataka State Open University  
Mukthagangotri, Mysuru-570006

### COURSE COORDINATOR

**Shilpa Rani N R**

Chairperson

Department of Studies in Library and Information Science  
Karnataka State Open University, Mukthagangotri, Mysuru-570006

### COURSE EDITORS

**Prof. M A Gopinath**

Professor (Retd.) in LISc  
DRTC, ISI Building, Mysore Road,  
Bangalore

**Prof. A Y Asuudi**

Professor (Retd.) in LISc  
Bangalore University  
Bangalore

**Dr. N. S Harinarayana**

Senior Lecturer  
Dept. of Library & Information Science  
University of Mysore, Mysore -06

**Prof. V. G. Talwar**

Professor in LISc  
Dept. of Library & Information Science  
University of Mysore, Mysore -06

### COURSE WRITER

### BLOCK EDITOR

**Dr. Devika P Madalli**

Associate Prof. Indian Statistical Institute  
DRTC 8<sup>th</sup> mile stone  
RV College PO, Mysore Road, Bangalore

### PUBLISHER

**Registrar**

Karnataka State Open University  
Mukthagangotri, Mysuru-570006

Developed by Academic Section KSOU, Mysore

**Copy Right: KARNATAKA STATE OPEN UNIVERSITY, 2017**

© All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from Karnataka State Open University.

This courseware is printed and published by The Registrar, Karnataka State Open University, Mysuru for limited use only. No individual or collaborative institution can use / print / distribute in any form without the written permission from KSOU. For user rights of this content and for other queries contact The Planning and Development Officer, KSOU, Mysuru 570 006.

Digital delivery of this courseware is also available for those who opt. For more details visit

[www.ksoustudymaterial.com](http://www.ksoustudymaterial.com) or [www.ksoumysore.edu.in/digitalcontent](http://www.ksoumysore.edu.in/digitalcontent)

## **M.Lib.I.Sc - 5: Information Systems: Architecture and Retrieval Block – 3 : Practical and Strategic Issues in Digitization of Library Collections**

---

### **Block Introduction**

Bill Gates the computer Wizard, bringing the analogy of online supermarket and home delivery in his address to the 1995 DLF (Digital Library Federation) Business Forum said Users will want their information accessible to them when they want it. Driving to a library which may not be open, hoping to find the information wanted is not the future model. The 24 hours reference library and IT services will supersede the check out circulation clerk.... The future can be utopian cybernetic or Orwellian in vision... we have a myriad of individual net accesses a plurality of Cyber villages but with most users ghettoized in their dedicated information and entertainment channels”. User becomes the focal point and he needs access to information and entertainment at his or her desktop in the office or at home”.

Developments in IT, the Internet and the Web have led to the development of the new era of electronic and digital libraries. Research in digital libraries is taking place in both developed and developing countries on account of rich funding by various national and international bodies. Digital libraries offer unique ways of recording, preserving and propagating culture in multimedia form. Apart from being organizations that preserve traditional culture (language, art, music, folk art etc) digital libraries serve to forward the frontiers of science. As Witten and Bainbridge opine “if information is the currency of the knowledge economy, digital libraries will be the banks where it is invested. Indeed Gothe once said that visiting a library was like entering the presence of great wealth which was silently paying untold dividends”.

Digital library architectures are very complex and designers tend to use their own terminology within the design process. The emphasis in architectural design should be user centric. Criteria for consideration while planning the architecture are 1) the services to be brought into the digitization, level of services, mode of delivery, interfacing and the

users' feedback and readjustment. General principles, in digital library architecture, its components and the essential features of the digital library software are examined in unit 9. The Greenstone software provides a convenient way of organizing information and making it available over the Internet. Making information available using this system is far more than just "putting it on the Web". The collection becomes maintainable, searchable and browser. The Greenstone software helps to meet these requirements.

Internet is bringing the world communities in one forum. This makes it imperative that resources in all worlds' languages to be transformed, indexed, searched and retrieved on Internet. There are three approaches to multilingual representation and retrieval on internet. These are 1). Character Encoding 2). Display and 3). Mapping for input. A character is a letter of an alphabet, a particular mark an a page, a symbol in certain language, and so on. Before the advent of the internet there were different computer architectures with different character sets and encodings. ASCII is an American Standard Code for Information Interchange. It is a character set using of 7 bit units with a trivial encoding designed for 7 bit bytes. ASCII is compatible with UNICODE without any specific change. UNICODE achieved the goal of representing the scripts of languages in use around the world. It is universal, in the sense that any document in an existing character set can be mapped into Unicode. The resulting Unicode file can be mapped back to the original character set without any loss of information, a feature called as roundtrip compatibility with existing coding schemes, and it is central to Unicode design. A most popular use of Unicode is multilingual content development over Internet. Since the Unicode provides a unique number for every character no matter what the platform is no matter what the language is and thus makes the web display easier. These are the issues addressed in unit 10 along with Unicode application.

Unit 11 explains the Human Computer Interaction. It is mainly focused on what are the types of users interactions with computers and the expectations of the users of how a system should allow for different levels and ways of interaction; the technology associated with it (hardware and software tools) and the guidelines for the presentation of information on the web.

Unit 12 discusses the issues in discovering the resources from the digital library collection. You have already studied the origins and structure of internet and the World Wide Web. The web is now the largest information space that the world has ever known and it continues to grow exponentially and it has a large quantity of content. As libraries are increasingly producers of digital information as well as consumers and organizers these resources are to be catalogued on-line. This is also important to help users find information in digital library environment. The purpose of creating tools for resource discovery is to allow users to locate the items they seek. Metadata is one of the principal concepts for the description, organization, exchange and retrieval of information in a networked environment. Metadata provide a means to discover what data set exists and how it might be obtained or accessed. Metadata help to document the content, quality, and features of a data set, indicating its fitness for use. Thus it is a process of substitution and a pointer to the location of the object. Dublin core is a metadata standard developed for describing resources on the web. The unit 12 describes its features. There are many kinds of metadata schemes. There are markup languages which allow metadata structures to be embedded within electronic text and can also be used to create metadata records for the description of both textual and non-textual resources. These are data description frameworks. Unit 12 presents in simple terms the resource discovery tools; the use of faceted classification for semantic organization of knowledge to enhance the access capabilities in networked environment.

**Prof. N B Pangannaya**

---

## **Unit – 9 : Architectural Agents and Tools for Digital Libraries**

---

### **Structure**

- 9.0 Objectives
- 9.1 Introduction
- 9.2 User Centric Design
- 9.3 Retrieval Issues
- 9.4 Digital Library Architecture
  - 9.4.1 Peer To Peer Architecture
  - 9.4.2 Grid Architectures
  - 9.4.3 Service-Oriented Architectures
- 9.5 Principles of Digital Library Architecture
- 9.6 Information Architecture
- 9.7 Components of Digital Library System
- 9.8 Digital Library Tools
- 9.9 Open Source Software for Digital Libraries
  - 9.9.1 Greenstone Digital Library Software (GSDL)
  - 9.9.2 E-Prints
  - 9.9.3 Fedora Digital Library
  - 9.9.4 D-space Digital Library
- 9.10 Summary
- 9.11 Answers to Self Check Exercises
- 9.12 Glossary of Key Terms
- 9.13 References and Further Readings

---

### **9.0 OBJECTIVES**

---

Reading this lesson will help to understand:

- Digital library architecture
- Principles of digital library architecture
- Components of digital library system
- Digital library tools

---

### **9.1 INTRODUCTION**

---

The proliferation of the World Wide Web and Internet technology has a significant impact on information service profession. Digital library is one direct result of the use and deployment of web for organized information services. However, the web itself is not a digital library. The web is an informal mechanism to connect documents in the form of web pages and files. Resource discovery is one of the biggest challenges in such vast and

unorganized collection such as the web. Contrary to this, Digital Libraries offer organized services with well-defined collections. They offer patrons services through various content managed modules and through resource discovery tools and standards that are implemented.

---

## 9.2 USER CENTRIC DESIGN

---

The central focus of an information system is of course the user requirements. Therefore the design of the system entirely depends upon the inputs from the users and users studies. There are various factors to be considered at stages in the design and development of the system to suit the user needs. To enlist a few:

- Scope
- Level of exposition
- Types of services
- Forms of document delivery
- Modes of dissemination
- Feedback incorporation
- User orientation
- User interfaces and interaction

The results of the user needs analysis culminate in providing user-friendly information products. The information services of a digital library take on a new perspective in the sense that the product is expected to give the information in a nutshell instead of a listing of documents in reply to an information query. Design of the services and implementation therefore form the crux of the work involved.

### a. Services

Information services of a digital library make it distinct in its operation and purpose. But a distributed environment has to cater to a more heterogeneous user community than a traditional library setup. This makes the planning and implementation of various services a challenging task. The applications for which the system will be used nor the level of performance deemed acceptable can be fixed.[1]. Further user needs and expectations develop along with the system, and their experience with the present and other systems will affect their expectations and evaluation. [2]. The main issues to be tackled in planning and implementation of the information services are:

**Scope:** The digital library should have distinct core of information according to the interest of the users and interest of the collaborating organizations. This may be supplemented with links to other sources such as Internet and other library databases for the service.

**Level of the services:** Usually the users in a digital library environment expect information and not the documents. The presentation is however important. It is to be

determined whether the information needs to be consolidated, qualified and whether other kinds manipulation into formats desired by the users can be provided. It is to be determined at what levels of users the information services are aimed. It may be to assist Policy makers, researchers, system developers and teaching faculty etc. A common system catering to a heterogeneous group may have to repackage the information to suit the varied requirements and power of assimilation.

**Mode of Delivery:** It is expected that most of the services be delivered on the desktop at the user's workplace. However the need for the documents in print may exist. So the system can have a sub-system, which may look into document delivery based upon the nature of the information need. One of the desirable features patron looks for in DLs, is subscription to collection. Subscription enables sending email alerts to subscribers holding them, whenever a new item is added to the collection.

**b. Interfaces**

Much emphasis is laid on design of user-friendly interfaces to system. Internet has made it possible to inter-connect all kinds systems and databases. This facilitates the designing of user interfaces to the digital library on the web using scripting languages. The most familiar clients available are the Internet browsers and act as the front-ends. However, each user has specific interest and follows distinct path of enquiry or interaction with the digital library. Hence, most digital libraries incorporate customization features to facilitate desired user interaction. Taking forward this further has led to research in personalization and visualization that study user needs and usability of DLs and their interfaces and based on experiences offer personalized services.

**c. Feedback and Re-adjustment**

The modules developed rely upon the user requirements. The fact that the users and their requirements vary, emphasizes that the system must be dynamic. The system designed should accommodate the feedback from the users and should be open to re-adjustment without the need for major over hauling.

**Self Check Exercises**

**1) What are user centric designs explain.**

Note: i. Write your answers in the space given below.

ii. Check your answers with the answer given at the end of this Unit.

---

---

---

---

---

---

---

---

---

### **9.3 RETRIEVAL ISSUES**

---

The success of the digital library of course lies in efficient retrieval. The system should have a flexible search language with online help at every stage. Further the system can be enhanced with the search options such as drop down menu of the subject strings from which the user may be able to choose from pre-coordinated strings. A set of example search queries and syntax may be provided to direct the users to formulate the search queries. The retrieved set should be available with abstracts so that they reflect the contents. Further links are provided to full text of documents to authorized users of digital libraries (depending of the access and rights policies).

---

### **9.4 DIGITAL LIBRARY ARCHITECTURE**

---

The digital library is often times described as librarianship shifted to the digital world. This description drives the argument that ‘physical’ and ‘digital’ collection is the only difference between the two. But digital libraries not only consist of transcribed physical objects, through processes such as digitization, but also of conceptual elements (improvisations) and digital objects that have no physical equivalents (web sites) [3]. Another distinct feature is that digital libraries encompasses not only collections but also services and various other components to enable a seamless access to organized collection in a distributed environment.

To encompass all the desired features and essential components, different digital library architectures have been put forward. The general trend is the shift from machine centric to user centric design. The underlying technology is a set of WWW clients communicating with httpd servers that use Common Gateway Interface (CGI) scripts and/or binaries to access a database [1].

The Kahn Wilensky architecture [4] provided fundamental aspects of an infrastructure that is open in its architecture and which supports a large and extensible class of distributed digital information services. The digital library elements considered, include digital objects, handles (unique identifiers of digital objects), metadata, repositories, handle generators, originators, users, global naming authorities and local naming authorities, and a repository access protocol (RAP).

However, Digital Libraries today have evolved as sophisticated systems, which are scalable, customizable and adaptive. Accordingly the architectural approaches are also following the trends in research for most community based information systems. The three approaches suggested by the DELOS project for DL architectures are as follows [5]:

#### **9.4.1 Peer to Peer Architecture**

Peer-to-peer (P2P) architectures allow for a loosely coupled integration of data and documents. Here autonomy between data and document is maintained and a mechanism for

collaborative sharing by clients systems is enabled. Different aspects of peer-to-peer systems are combined and integrated into an infrastructure for digital libraries.

#### **9.4.2 Grid Architectures**

Grid architecture follows the idea of a service grid that includes handling of shared resources. Grid architecture for digital libraries aims at integrating functional and service components to build a suitable infrastructure.

#### **9.4.3 Service-oriented Architectures**

This architecture aims at providing common service interfaces based on the standards adopted by digital libraries. Not only digital objects but services have also to be semantically represented so that they can be discovered by the users. This architecture incorporates service description languages stored in service registries.

#### **Self Check Exercises**

**2) Mention the different approaches as suggested by the DELOS project for DL architectures?**

**3) What is grid architecture?**

Note: i. Write your answers in the space given below.

ii. Check your answers with the answer given at the end of this Unit.

---

---

---

---

---

---

---

---

---

---

### **9.5 PRINCIPLES OF DIGITAL LIBRARY ARCHITECTURE**

The Digital library trends show it is not only technology and technological advances that have affected how digital libraries take shape. Legal and social implications also have an equal impact. The general principles guiding Digital Library architecture stated by William Arms provide a holistic approach [6]:

1. The technical framework for digital libraries exists within a legal and social framework
2. Understanding of digital library concepts is hampered by terminology
3. The underlying architecture should be separate from the content stored in the library
4. Names and identifiers are the basic building block for the digital library
5. Digital library objects are more than collections of bits

6. The digital library object that is used is different from the stored object
7. Repositories must look after the information they hold
8. Users want intellectual works, not digital objects

Digital libraries evolved from web based services and are often treated as extensions of collections of web resources. But the academic world today realizes digital libraries encompass much more than what the early networked information systems offered. The functions follow different workflows for each digital library and its environmental and compositional factors. Some generalizations are made to give a framework for digital library architecture that serve as generic models. The Dienst architecture follows the following principles [7]:

**•Open Architecture**

Essentially involves functionality partitioned into set of well-defined service and services accessible via well-defined protocol. Well-defined services with protocols to enable federation and interoperability

**•Modularization**

Modular architecture promotes interoperability and the products are scalable to different clientele. Modular designs increase flexibility and extensibility

**•Federation**

Enables aggregations into logical collections and access to distributed collections through a single interface.

**•Distribution**

Distribution of content (collections) and services and facilitate distributed administration and management of DL

The Dienst system encompasses:

- Document model
- Naming service (CNRI's Handle System)
- Repository service
- Indexer service
- Collection service
- User Interface service

The federated approach has given way to centralized approach in resource discovery. The open archives initiatives architecture has put forth the idea of Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) that facilitates the concept of Open Digital Libraries. In this model Digital libraries act as data providers and service providers. Compliance with this standard implies the data provided (mostly metadata) is harvested and

stored for retrieval purposes so that search is not federated in real time but results can be given instantaneously.

### Self Check Exercises

#### 4) List out the general principles guiding Digital Library architecture.

Note: i. Write your answers in the space given below.

ii. Check your answers with the answer given at the end of this Unit.

---

---

---

---

---

---

---

---

---

---

---

## 9.6 INFORMATION ARCHITECTURE

---

Information architecture within digital libraries pertains to the design and structure of the collections as presented to the DL patrons. This essentially relates mostly to the taxonomy of the topical domains familiar to the clientele and DL environment. In the words of Arms, etal, [8], the information architecture is guided by the following basic principles:

- Users and their applications programs must be given flexibility. Since users explore material in almost every conceivable manner, the organization of information should not be biased by expectations about how users will approach the material, their level of expertise, or the sequence in which items will be accessed.
- Collections must be straightforward to manage. In digital libraries, as in all libraries, comparatively small professional staffs manage very large collections of material. The architecture must allow the staff to concentrate on curatorial aspects, and free them from routine tasks wherever possible.
- The information architecture must reflect the economic, social, and legal frameworks developing in the information infrastructure. In particular it must recognize that information is valuable, subject to terms and conditions, and is transmitted over insecure networks that cross national boundaries.

To illustrate information architecture in one of the popular DL tools consider the DSpace Digital Library. Here the digital library collections are divided into communities and sub-communities. The communities and sub-communities contain collections. Typically the communities are by departments or research labs etc and each distinct community will have its respective collections of items. This facility ensures hierarchical organization of the content of digital libraries thereby providing at a glance the scope of the collections and

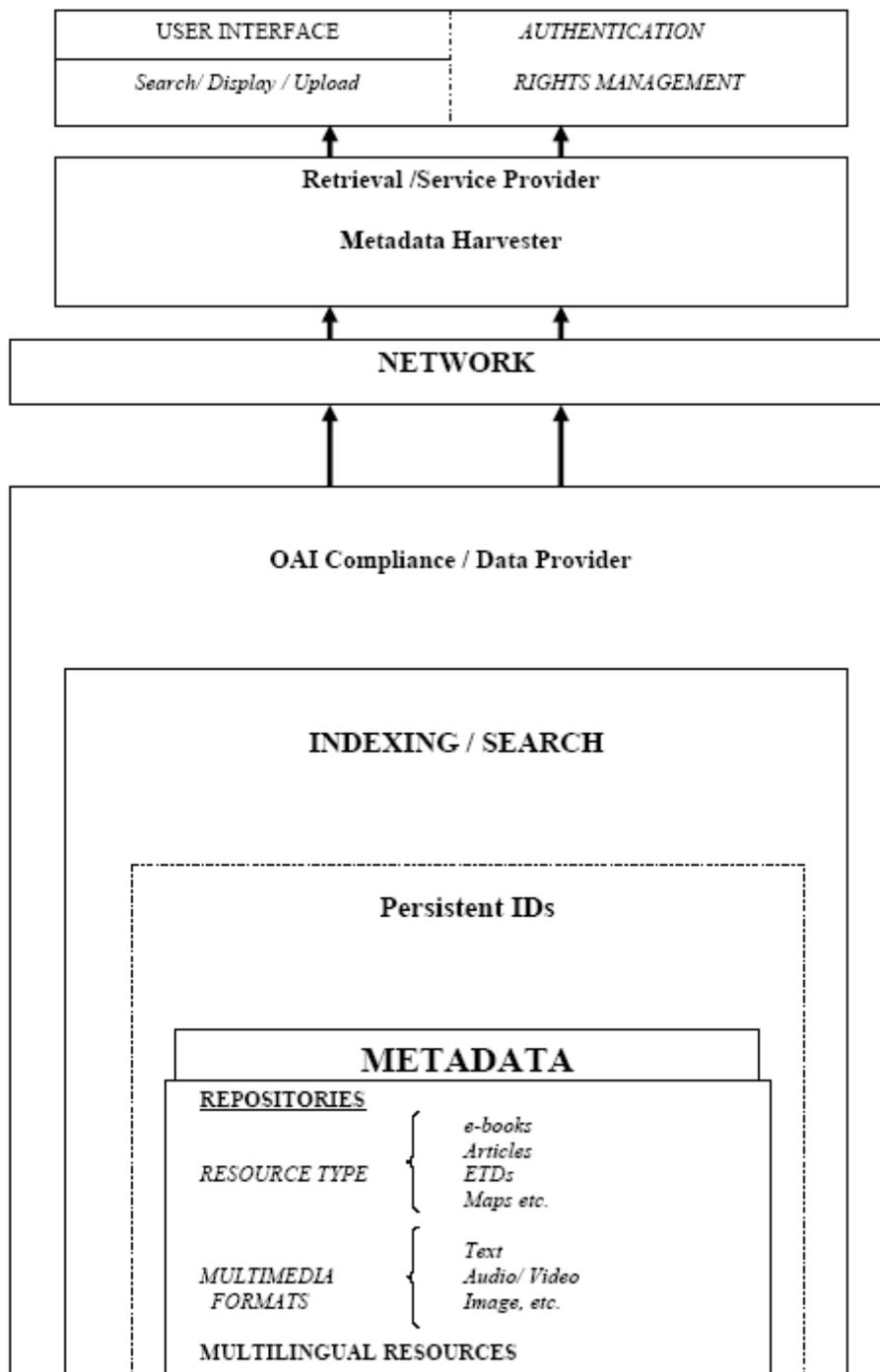




initiative has to first deliberate upon, sort out and plan for with unambiguous policy support.

Once the policy decisions are made regarding the content, authors, etc, other factors such as choice of the type of system to adapted, technological and other criteria should be drawn up. The expected features of digital library software can be enlisted as follows [10]:

1. *Low cost*, (including all hardware and software components)
2. *Technically simple* to install and manage;
3. *Robust*
4. *Scalable*
5. *Open and inter-operable*
6. *Modular*
7. *User Friendly*
8. *Multi-user* (including both searching and maintenance)
9. *Multimedia* digital object enabled; and
10. *Platform independent* (including both client and server components).



*Fig 1: Components of Digital Libraries*



Greenstone Digital library has a windows based version along with UNIX and MAC/OS platforms. Hence it is popular among libraries that use MS-Windows platform. It has documentation in English, French, Spanish and Russian languages [11]. It has a flexible search facility that enables searching by text. Searching can be selective also by paragraphs and sections.

GSDL has multilingual interfaces in the languages Arabic, Chinese, Czech, Dutch, French, Galician, German, Hebrew, Indonesian, Italian, Japanese, Kazakh, Maori, Portuguese, Russian, Spanish, Thai, Turkish and English.

GSDL can be customized for its appearance and other functions like searching and browsing. Searching can be by title or authors. For the subject approach a classificatory hierarchy can be included. GSDL handles various document formats but it can be specified which all document formats will be included in the collections such as text, PDF, HTML etc. also the look and feel of the pages in terms of the colors, headers, logos and notes can be customized as required. Display of search results can also be customized.

### **9.9.2 EPrints**

GNU EPrints is generic archive software by the University of Southampton. GNU EPrints primary goal is to be set up as “an open archive for research papers, and also other objects such as images, research data, audio archives - anything that can be stored digitally, with some changes to the configuration” [12]. Eprints has most installations up until now in academic repositories. It is highly configurable for collections, metadata schema and their content. The EPrints software has been developed under GNU/Linux and is intended to work on any GNU system. It works on the any of Unices/Solaris while some installations are on MAC (OS/X). Copy of Eprints digital library software can be downloaded from <http://software.eprints.org/download.php>.

Eprints has a standard installation procedure by the configure procedure. It has a modular structure and tools and scripts can be added as required [11]. Eprints allows any of the known document formats. An interesting feature is that a single document can be stored in more than one format. Eprints allows use of any metadata schema. The DL can have its set of elements and for each e-print type. Mandatory and optional fields for metadata can be specified. Authors also can have associated metadata. EPrints is compliant with OAI PMH standard for interoperability.

It has powerful search mechanism. One interesting feature is searching by subject hierarchy. The standard search approaches like free text, Boolean are also provided.

Submission is through web interface. At the time of submission the contributor can use a subject hierarchy under which object will be indexed. Submissions can be single files or zipped files or websites also. By subscribing to collection users can get mail alerts about the new additions to collections.

EPrints has an inbuilt review process. Hence repository can have in place a reviewing panel and workflows can be set accordingly. The moderation process is also through web interface.

### **9.9.3 Fedora Digital Library**

The Fedora project funded by the Andrew W. Mellon Foundation to build an open-source digital object repository management system based on the Flexible Extensible Digital Object and Repository Architecture (Fedora). Fedora was developed as a research project at Cornell University, and successfully implemented at University of Virginia as a prototype system to provide management and access to a diverse set of digital collections.

“Fedora is a general-purpose digital object repository system that can be used in whole or part to support a variety of use cases including: institutional repositories, digital libraries, content management, digital asset management, scholarly publishing, and digital preservation”[11]. The Fedora repository system is open source software licensed under the Mozilla Public License. Fedora DL software is available for download from <http://www.fedora.info/>.

Fedora is an integrated system dealing with resources, metadata and other attributes of an object through the Metadata Encoding and Transmission Standard (METS). Each resource is a fedora digital object and each object is container for Data Streams. A DataStream maybe an object representing content or metadata. Further the DataStream may reside inside fedora repository or may be linked and pointed out by referencing [13]. Fedora is compliant with the OAI-PMH protocol for interoperability. A good feature of fedora is versioning where various version of an object can be stored and retrieved. This feature is of special importance for systems dealing with preservation of digital objects.

### **9.9.4 DSpace Digital Library**

DSpace was initially developed by Hewlett-Packard and MIT in collaboration with the objective to create a package that could provide an institutional repository, which addressed the problem of digital preservation as a central theme. A number of individuals from institutions using DSpace have taken on the role of developers, and a community of interested parties has evolved who have started to feed code back into the core. DSpace future development and implementation is part of the open-source development model.

The application provides ways of capturing, storing, indexing, preserving and disseminating digital objects. The collections can include, research papers, conference papers, book chapters, datasets, learning objects and, of course, E-Theses.

DSpace facilitates the following [14]:

- Captures – Digital research material in any formats directly from creators
- Describes – metadata –Descriptive, technical, rights metadata
  - Persistent identifiers – using the CNRI handle system

- Distributes – provides search and delivery of resources through web
- Access control and Authentication
- Preserves - Large-scale, stable, managed long-term storage – archival and preservation

DSpace has incorporated features required for sophisticated and integrated operations of digital libraries. Accordingly its functions include comprehensively all activities from collection planning, building and web based information services.

The information model may be planned according to organizational structure and Dspace facilitates organizations of collections accordingly. Each DSpace repository maybe divided into *communities* at the highest level. Communities typically correspond to a laboratory, research center or department. Typically in an academic institution it is by the various departments or faculties of study. As of DSpace version 1.2, these communities can be organized into a hierarchy. Communities contain *collections*. Each collection further comprises *items*. Items are the basic archival elements. Each item is owned by one collection though it may be mapped to others. Items are further divided into bitstreams that are computer files.

**Metadata:**

Metadata is of three types in DSpace:

*Descriptive:*

Descriptive Metadata: DSpace uses qualified Dublin Core metadata elements and provides worksheets in the submission process. It allows the DL manager to select the subset of elements as per requirements or to add more elements.

*Administrative:*

This includes preservation metadata, provenance and authorization policy data.

*Structural:*

This includes information about how to present an item, or bitstreams within an item, to an end-user, and the relationships between constituent parts of the item.

Users, Members and DSpace managers are managed through the e-people and authorization modules. This module deals with members and their authorizations. Users of a DSpace repository may browse and search collections. They may be authorized as authors or “Submitters”. Additionally administrative and reviewing responsibility may also be assigned to some of the members. Further users can be grouped and group authorizations can be managed which is an ideal way to manage department or faculty wise categorized collections and activities.

In the open access repositories to ensure quality of the content a strict review process should be put in place so that material in the open access will gain the same credibility as

that of “ranked journals”. DSpace allows defining a workflow for the reviewing process. The workflow is setup for each collection. Reviewers can be assigned for reviewing the contents as well for correcting the metadata of the items.

Browsing facility in DSpace allows the users to browse by communities and collections (hierarchies are indicated in the community homepage). Users may also browse by the authors and titles (alphabetical index) and by date across collections or within a particular collection. DSpace incorporates Jakarta’s Lucene search engine. In addition to normal search features such as Boolean, wildcard searches, it has advanced features of searching such as the Levenshtein Distance algorithm for fuzzy logic and term boosting (assigning weights to words in a query).

In DSpace metadata is XML tagged for all transactions such as harvesting and migrating of records. DSpace is OAI version 2 compliant and allows harvesting the records in XML carrying the DC elements.

One of the valid apprehensions of patrons of online repositories is that the resources are volatile. They tend to disappear or change without any notice. Therefore it is important to give persistent identifiers to the digital objects. DSpace uses Handles primarily as a means of assigning globally unique identifiers to objects. DSpace uses the CNRI Handle System for creating these identifiers.

Another interesting feature of DSpace is support for OpenURL. OpenURL enables opening an item from a referred link to it. DSpace displays an OpenURL link on every item page, automatically using the Dublin Core metadata.

End-users (e-people) may subscribe to collections. They receive an email alert when new items appear in those collections. They may unsubscribe any time they choose.

Records can be imported into and exported from DSpace repositories. DSpace includes batch tools to import and export items in a simple directory structure, where the Dublin Core metadata is stored in an XML file.

---

## **9.10 Summary**

---

Digital libraries are integration of theory and practice of librarianship, online-networked resources, various media and standards and protocols required for transactions. The scenario poses interesting challenges to computer professionals and Digital Libraries have attracted their interest. Open source voluntary efforts in the development of DL tools have offered very efficient open source software for building DLs. Each is unique by its features as discussed in the sections above. The choice of system basically depends upon the type of resources, the size of data, user community and its needs and the kind of information services expected of the Digital Library.

---

## 9.11 ANSWERS TO SELF CHECK EXERCISES

---

- 1) The central focus of an information system is of course the user requirements. Therefore the design of the system entirely depends upon the inputs from the users and users studies. There are various factors to be considered at stages in the design and development of the system to suit the user needs. To enlist a few:
  - Scope
  - Level of exposition
  - Types of services
  - Forms of document delivery
  - Modes of dissemination
  - Feedback incorporation
  - User orientation
  - User interfaces and interaction
- 2) The three approaches suggested by the DELOS project for DL architectures are as follows:
  - Peer to Peer Architecture
  - Grid Architectures
  - Service-oriented Architectures
- 3) Grid architecture follows the idea of a service grid that includes handling of shared resources. Grid architecture for digital libraries aims at integrating functional and service components to build a suitable infrastructure.
- 4) The general principles guiding Digital Library architecture as stated by William Arms are as follows:
  1. The technical framework for digital libraries exists within a legal and social framework
  2. Understanding of digital library concepts is hampered by terminology
  3. The underlying architecture should be separate from the content stored in the library
  4. Names and identifiers are the basic building block for the digital library
  5. Digital library objects are more than collections of bits
  6. The digital library object that is used is different from the stored object
  7. Repositories must look after the information they hold
  8. Users want intellectual works, not digital objects
- 5) According to Arms, etal, the information architecture is guided by the following basic principles:

- Users and their applications programs must be given flexibility. Since users explore material in almost every conceivable manner, the organization of information should not be biased by expectations about how users will approach the material, their level of expertise, or the sequence in which items will be accessed.
  - Collections must be straightforward to manage. In digital libraries, as in all libraries, comparatively small professional staffs manage very large collections of material. The architecture must allow the staff to concentrate on curatorial aspects, and free them from routine tasks wherever possible.
  - The information architecture must reflect the economic, social, and legal frameworks developing in the information infrastructure. In particular it must recognize that information is valuable, subject to terms and conditions, and is transmitted over insecure networks that cross national boundaries.
- 6) Digital libraries comprise of components that handle repositories building, maintenance, searching and retrieval and distributed information services in a networked environment. The components of digital libraries include the following:
- Resources/Collection
    - Types of resources
    - Formats
    - Multilingual Collection
  - User interface
    - Memberships/login
    - Submission
    - Search/save
    - Alert system
  - Metadata Standard and Interface
  - Indexing
  - Search/Browse/Display facility
  - Interoperability Standards
  - Import/Export modules
  - Persistent Identifiers
  - Authentication and Rights/Access Management
  - Network elements
- 7) The expected features of digital library software can be enlisted as follows:
- Low cost, (including all hardware and software components)
  - Technically simple to install and manage;
  - Robust
  - Scalable
  - Open and inter-operable
  - Modular

- User Friendly
- Multi-user (including both searching and maintenance)
- Multimedia digital object enabled; and
- Platform independent (including both client and server components).

---

## 9.12 GLOSSARY OF KEY TERMS

---

**Metadata:** Data that is used to describe other data.

**Dublin Core:** It is a metadata standard for describing digital objects (including webpages) to enhance visibility, accessibility and interoperability.

**Open Source:** A movement offers products that can be used, modified and adapted to provide solutions in various systems without or at a nominal cost.

**OAI-PMH:** Open Archive Initiative-Protocol for Metadata Harvesting.

**GSDL:** Greenstone Digital Library Software.

**EPrints:** GNU EPrints is generic archive software by the University of Southampton.

**DSpace Digital Library:** Dspace developed by Hewlett-Packard and MIT in collaboration with the objective to create a package that could provide an institutional repository.

**Fedora Digital Library:** It is an open-source digital object repository management system based on the Flexible Extensible Digital Object and Repository Architecture (Fedora).

---

## 9.13 REFERENCES AND FURTHER READINGS

---

1. Berners-Lee, T. J., Cailliau R., Groff, J. F., Pollermann B. 1992. World-Wide Web: The information universe. Electronic Networking: Research, Applications and Policy 2(1) (Spring), pp. 52-58.
2. PARASURAMAN (A), BERRY (Leonardo L.) AND ZEITHAML (Valarie A.). An empirical examination of relationships in an extended service quality model. Marketing Science Institute Working Paper, Report 90-122, Cambridge, MA,1990.
3. Digital libraries: issues and architectures.  
<http://www.cSDL.tamu.edu/DL95/papers/nuernberg/nuernberg.html> (browsed on 22/09/2005)
4. Robert Kahn and Robert Wilensky. A Framework for Distributed Digital Object Services,<http://www.cnri.reston.va.us/home/cstr/arch/k-w.html> (browsed on 25/09/2005)

5. Digital Library Architecture – DELOS site. [http://ii.uit.at/research/delos\\_website](http://ii.uit.at/research/delos_website) (browsed on 22/09/2005)
6. William Y. Arms, Key Concepts in the Architecture of the Digital Library. D-Lib Magazine, July 1995. <http://www.dlib.org/dlib/July95/07arms.html> (browsed on 24/09/2005)
7. Sandra Payette. Digital Library Architecture: A Service-Based Approach. <http://www2.cs.cornell.edu/payette/presentations/DL-architecture.ppt> (browsed on google on 25/09/2005)
8. William Y. Arms, Christophe Blanchi and Edward A. Overly. An Architecture for Information in Digital Libraries. D-Lib Magazine, February 1997. <http://www.dlib.org/dlib/february97/cnri/02arms1.html#info-arch> (browsed on 24/09/2005)
9. Berring, RC 1993, 'Future librarians', in Future libraries, RH Bloch and C Hesse, eds. Berkeley, University of California Press pp. 94-115.
10. Pandey, Richa. Digital Library Architecture. In Workshop on Digital Libraries: Theory and Practice. March 2003. DRTC, Bangalore.
11. Greenstone Digital Library Software. <http://www.greenstone.org/cgi-bin/library>.
12. GNU EPrints Documentation- Introduction. <http://software.eprints.org/docs/php/intro.php> (browsed on 24/09/2005)
13. Fedora Digital Library Documentation. <http://www.fedora.info/documentation/> (browsed on 24/09/2005)
14. DSpace System Documentation. <http://www.dspace.org/technology/system-docs/> (browsed on 24/09/2005)

---

## **Unit – 10 Multilingual and Multi script issues**

---

### **Structure**

- 10.0 Objectives
- 10.1 Introduction
- 10.2 Multilingual Resources
  - 10.2.1 Character Encoding
  - 10.2.2 Display Issues
  - 10.2.3 Mapping for Input
- 10.3 UNICODE Standards
- 10.4 UNICODE and its Features
- 10.5 Web Content Development Through Unicode
- 10.6 Applications of Unicode
- 10.7 Summary
- 10.8 Answers to Self Check Exercises
- 10.9 Glossary of Key Terms
- 10.10 References and Further Readings

---

### **10.0 OBJECTIVES**

---

This unit will help to understand:

- Basics of character encoding
- Multilingual representation
- Standards for character encoding
  - ASCII
  - ISCII
  - UNICODE
- Challenges in Multilingual and Cross-lingual retrieval.

---

### **10.1 INTRODUCTION**

---

Transactions on computers and networks for most part are in English. With the Internet bringing the world communities on one forum there is a need for resources in all worlds' languages. In an effort to give information services through Digital Libraries tackling multi-lingual resources becomes a crucial issue. There are many facets to be considered from representation of multilingual data to providing indexing, searching, and retrieval. There is a need to study how characters are represented within computers and how they are rendered. One of the first methods used for giving regional language was by using some proprietary technology and fonts to display different languages. However, these were not

in any standard encoding form and hence such data could not be processed for searching and retrieval by other tools.

---

## **10.2 MULTILINGUAL RESOURCES**

---

The older techniques available for representation of other languages are mostly product specific. Each product had its own internal representation or rendering form that could not be interpreted across applications. These were very much dependent upon particular tools. However, complex procedures are required for multi-lingual information retrieval wherein the following two important functions need to be supported: 1) supporting documents and queries in several natural languages simultaneously (multilingual retrieval) and 2) making it possible for queries in one language to retrieve related documents in other languages (cross-lingual retrieval)[1]. For such operations merely concentrating on displaying information in different languages does not suffice. It calls for a universal encoding system that would allow access and processing of the information globally.

Mainly there are three aspects to multilingual representation and retrieval on Internet.

- Character Encoding
- Display
- Mapping for input

### **10.2.1 Character Encoding**

Computers understand all data in terms of 1s and 0s, which is basically a binary representation system. So whatever alphabets or numerals are fed into the computers have to be finally transformed to the respective binary form to be understood, stored or processed. Also, to display a character from whatever language, that character needs to be represented in some way for computers to be able to understand it. The way characters are 'encoded' internally is called a *character encoding*. There are many different kinds, each using different numbers for encoding characters. One popular code which has long been in use is ASCII .

#### **10.2.2.1 ASCII**

**ASCII** stands for "American Standard Code for Information Interchange" and has been the standard for textual data processing and transfer. ASCII is a 7-bit binary code for representing characters. That is using ASCII, 128 (0-127) characters could be represented. Most data input into computers is done through the standard QWERTY keyboards. ASCII assigns a value for each keystroke in terms of a number which is finally stored in a 7-bit binary number. For example, the letter A is assigned the number 65 and the binary form of it 1000001 is stored whenever A is entered. Likewise the values for alphabets, numerals and special characters like form feed and line break are assigned numbers. But only English characters are included in this and that is why most computers communicated in English as the standard allowed only for it.

Character on the screen	Binary value used to process it	Character on the screen	Binary value used to process it
1	0110001	A	1000001
2	0110010	B	1000010
3	0110011	C	1000011
4	0110100	D	1000100
5	0110101	E	1000101

Thus when we consider languages other than English, we start wondering how the characters of these should be encoded. Extended ASCII which is a 8-bit code allows for 256 character representations and this is often used to represent other languages from position 128 to 255.

### 10.2.2 Display Issues

Codes like ASCII dictate how the characters should be stored (memory representations). But how to display them on the screen or send to the printer is another matter. This completely depends on fonts available and supported by the operating system and application software. Each character is displayed through a 'glyph'. Glyphs represent the shapes that characters can have when they are displayed. Glyphs are images. The final appearance of the rendered text can be often quite different from the shapes of the individual characters that constitute the textual representation. The character to glyph rendering mechanism is the domain of software processes [2]. Although the encoding encapsulates only the basic alphabetic characters, the number of glyphs and their combinations required for the exhaustive rendering for scripts like Indic scripts can be quite large. A set of glyphs for a script used to display text constitutes a font. The number of glyphs can range from a couple of hundred to a couple of thousand or more depending on the complexity of the script and the font design. Computer systems have resident fonts (mostly English) that make text visible. So if the character A is to be displayed as bold or italics or as particular type face such as Arial or Times New Roman, then the operating system (like Windows 2000) and the application software displaying it, should support it through the necessary typefaces.

### 10.2.3 Mapping for Input

There should be some mechanism of inputting data, the most common being the keyboards. The most common language seen on the keys of keyboards are the English alphabet (mostly the keyboards are referred as QWERTY keyboards). The keystrokes are mapped so as give the internal coding for each alphabet keyed in. What about other languages? There are a few customized keyboards for different languages like Hindi or Kannada and other Indian languages. European countries also use keyboards with keys for their respective scripts. They may be available physically on the keyboard. Or else on the computer screen a soft keyboard may be available called as 'virtual keyboards'. These can enable one language at a time because the keyboard strokes are to be mapped uniquely, for the encoding sets of characters. English set as pointed out, is universally accepted and available. For other

languages such as Indian languages there should be support from the OS and application software for input, encoding and display.

### Self Check Exercise

1) What are the different aspects to multilingual representation and retrieval on Internet?

2) What is ASCII?

Note: i. Write your answers in the space given below.

ii. Check your answers with the answer given at the end of this Unit.

---

---

---

---

---

---

---

---

---

## 10.3 UNICODE STANDARD

---

UNICODE is developed by UNICODE consortium. This consortium is a group of software organizations like IBM, Microsoft, Xerox, Oracle etc. and they came out with a 16-bit code and called it UNICODE, as it promises to cover all the world's scripts. The promising feature of UNICODE is, it can represent 65,536 characters. The first version of UNICODE came in 1991.

It is fully compatible with ASCII i.e. the first 128 characters are ASCII characters. Thus any software using ASCII is compatible with UNICODE without any specific changes. UNICODE has defined different language zones for their corresponding scripts. For example, Devanagari characters start at 2304 and ends at 2431. Thus it includes all the scripts of the world.

### 10.3.1 Design principles of UNICODE:

- It is a 16-bit code
- For future expansion surrogate area has been defined which can generate combinations of two 2-bytes packets.
- UNICODE bothers about characters not glyphs.
- It renders logical ordering for writing i.e. from left to right and in case of Hebrew, Urdu etc. right to left.

- Combining characters also follow logical order. Combining characters are non-spacing characters, which form combination to represent a new character.

### 10.3.2 Structure of UNICODE

UNICODE initially was the interest of software organizations. Earlier, to produce a different language software, it was required to use ASCII code in the same language. But once UNICODE came up it was really not needed to rewrite the software for different language. So it has become a boon for software developing organizations side-by-side it became popular as it supports multilingual communication, which ASCII failed to do.

UNICODE is fully compatible to ISO/IEC 10646 set of standard developed by International Organization for Standardization and International Electrotechnical Commission in 1993. This another very interesting venture, which also promises 16-bit code, but it does not restrict to 16-bits. It promises to give even 32-bit code, if it really required for character representation. The use of term multiple octet is used because multiples of 8-bits are used in system. This code is called as Universal Multiple-Octet Code (UCS). Since this character code system uses 16-bit and 32-bit, it is known as UCS-2 and UCS-4 respectively.

### Self Check Exercises

#### 3) What is UNICODE? Mention the designing principles of UNICODE.

Note: i. Write your answer in the space given below.

ii. Check your answer with the answer given at the end of this Unit.

---

---

---

---

---

---

---

---

---

---

---

## 10.4 UNICODE AND ITS FEATURES

---

UNICODE is a 16-bit character code and it is an effort to represent all the existing word scripts. UNICODE mainly has to deal with the following [3]:

- It has to take care of existing ASCII coding system
- There are characters, which use a combination of characters, so such combinations have to be managed. (Strictly speaking this problem has not been solved yet, because it is the property of the fonts which are used in writing UNICODE characters)

#### **10.4.1 Sixteen bit character code**

It is known that UNICODE is 16-bit character code. It can accommodate 65,536 characters. But consortium has not exhausted all the character positions, 2048 character positions are left for the future characters. This region of UNICODE is known as Surrogate area. Surrogate area covers 55,296-57,343 character positions.

#### **10.4.2 Efficiency**

UNICODE is no doubt a very efficient code. All like characters are not repeated. They can be simply borrowed from some other group where they actually occur. For example, Hindi characters are from 2304 to 2431, Bengali is from 2432 to 2559, telugu is from 3072 to 3199 and so on.

For Carriage Return (CR) the code value is 14 (good old ASCII value) and is retained in UNICODE along with English alphabets. While encoding other scripts where CR is required, UNICODE still uses the value 14. This is uniformly applied to all the special characters of ASCII. Thus UNICODE avoids the repetition and adds efficiency to the coding system.

#### **10.4.3 Characters not glyphs**

In a sentence the smallest unit is a character, which can be written and can have semantic value. Glyphs are nothing but exposition of characters in graphical form, i.e. how a character appears on display, for example, bold, italics, or whatever. Glyphs are decided by the fonts and the actual value of UNICODE character will remain the same. To display the character glyph, it is necessary that a font must have a glyph in its source, that is why, it is said that it is the property of font used for writing.

If one tries to write in the word 'JAVA<sup>TM</sup>', the appearance of TM on the head of JAVA is due to glyph which shows superscript and is supported by the font. But the value of T will be 84 and M will be 77. That is why it is said that UNICODE is a code for characters not for glyphs.

#### **10.4.4 Plain text**

As it has been seen in the last section that UNICODE has nothing to do with glyphs, that means it only works out the plain text. For example, if a character is written '65' it is represented as 'A'(which has value in binary system 0000000001000001) irrespective of the format added with the character, which is known as 'Rich character'.

#### **10.4.5 Logical order**

UNICODE stores characters in logical order. Logical order means the order characters are stored in memory is same the order characters are keyed in using the keyboard. At certain places it has to be violated because the representation of the text is reverse. For example, Urdu.

Urdu is written from right to left and same is the case with Hebrew too. So when a character is typed it displays from right to left, but the order in which the character is represented in the memory is same. And UNICODE does it automatically. This particular feature of UNICODE is known as bi-directional ordering.

#### 10.4.6 Unification

UNICODE has taken care to identify the same character within different language and put it at one place in UNICODE table. For example, Punctuation marks, common letter etc.

The characters of East Asian scripts are taken in a separate table called as Han table. It covers Japanese, Korean, Chinese characters because these scripts have lot of overlapping characters, many a time the same character is used in all languages. Similarly, UNICODE has 'Combining characters table' which covers non-spacing characters that can be combined with other characters.

#### 10.4.7 Compatibility characters

Compatibility characters are the characters, which actually would not have been encoded because these are the variant form of some characters, which are already encoded.

For example,

Colon (:) {character position 58} has two Compatibility characters one at character position 847 i.e. 'Ratio' and other is 1417 i.e. Armenian full stop. There is only formatting variance between Colon and Ratio. In 'ratio' the distance between the points is more.

These characters reside in Compatibility area as one can see the index of character positions, but it is not necessarily true that all characters reside in this area only. For example, the variant of Latin character 'A' (character position 65) has another variant Full length Latin character 'A' which is given a position at the end of UNICODE table.

#### 10.4.8 Dynamic composition

Dynamic composition is used to generate different accent forms. There is a block in UNICODE character set map called as 'Combining characters'.

To generate, À, the characters A and ` are used.

À → A + `

The glyph À is permitted in the font set used. In Indic languages, both vowels and consonants can be used as Combining characters (non-spacing characters).

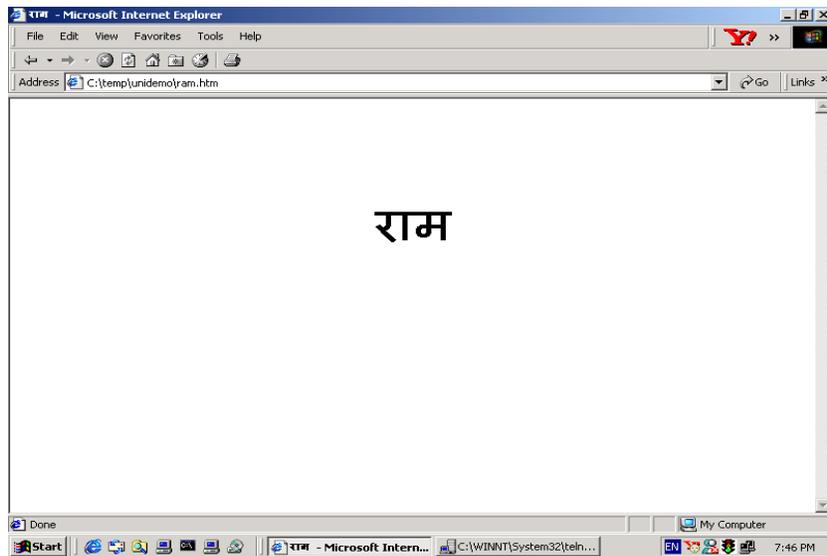
#### 10.4.9 Equivalent sequence

These are pre-imposed character codes available in the UNICODE character set. But same characters may be generated using composed characters. For example,

À → A + `

192                      65                      768





**Fig. 2:** Content Representation in Devnagari

Code using Devanagari script in Windows 2000

```
<html>
<title>राम</title>
<body><center><font size="7">
राम</font></center>
</body>
</html>
```

HTML page using the UNICODE values in Windows 2000

```
<html>
<title>&#2352&#2366&#2350</title>
<body><center><font size="7">
&#2352&#2366&#2350</font></center>
</body>
</html>
```

Both approaches are good but since Windows 2000 gives the facility of virtual keys to type Devanagari fonts, it is easier. Using UNICODE facilitates writing in more than one script in a single document.

---

## 10.6 APPLICATIONS OF UNICODE

---

UNICODE promises to cover all the world scripts including extinct scripts as well.

### **10.6.1 Multilingual encoding**

Most popular use of UNICODE is in multilingual editing. One can write in any language, if he/she knows the language and codes. But the problem of such writing is to know codes for all characters. So definitely some multilingual editor is needed, with which we can write in any language. The advantage of it is that each script can be read in any UNICODE compatible editor.

### **10.6.2 Multilingual transliteration**

Once being aware of the character codes, the approach towards transliteration can be taken. This feature is particularly advantageous with regard to Indian languages because all Indian languages use more or less the same consonants and vowels. Though some languages have some additional characters, in UNICODE these are arranged in such a way that there is fixed difference in the value of same phonetic character belonging to different scripts. For example, the difference between all Telugu and Devanagari characters is 768.

### **10.6.3 Applying UNICODE to the libraries**

India is a country of diversity. Many languages are spoken and many scripts are used. According to one survey, English is the language spoken by less than 10%. Besides, India holds voluminous information in many of the native languages. Libraries particularly Digital libraries can record information in the language and script of the source document. Even bibliographic information can be represented in the language of the document. Once UNICODE is adopted, one can exploit the advantages of transliteration so that data can be transliterated wherever and whenever it is required. It is not uncommon that much of the Sanskrit literature is available in different scripts of India. In such cases, even if Sanskrit text is keyed in using Devanagari, one can read the same text in his script (for example Telugu). In other words, it is not required to enter the text in Telugu, rather the Devanagari script can be transliterated on the fly.

In case of metadata (bibliographic information), the semantics of information can be given using XML (Extensible Markup Language), provided that libraries participating in the program should have consensus on a definite set of tags. Thus catalogue of one library can be seen and read in other language [4].

### **10.6.4 Indic Scripts**

Indic scripts are phonetic in nature. There are vowels and consonant symbols. The consonants become a syllable after the addition of a vowel sound to it. The point to be noted is that after the addition of the vowel, the original consonant symbol is modified to show the addition and is not as simple as adding the vowel next to it as is the case in English. Further complication for representation is that there are '*compound syllables*' also referred as ligatures. For instance, consider 'tri' in 'triangle' – here there are three letters corresponding to three sounds 'ta', 'ra', 'yi'. But in the case of Indic Scripts the three are built together to make a single compound consonant having a non-linear structure. (Illustrated below:)

**Illustration:**

**English** - **ta + ra + yi** → **tri**  
**Hindi** - **त + ष् + र + ि** → **त्रि**  
**Kannada** - **ತ + ಷ್ + ರ + ಿ** → **ತ್ರಿ**

Fig1: demonstrates ligatures in Indian Languages

**Display:**

The main problem with display of Indic scripts is dealing with their non-linear structures. Glyphs have variable widths and have positional attributes. Vowel signs can be attached to the top, bottom, left and right sides of the base consonant. Vowel signs may also combine with consonants to form independent glyphs. Consonants frequently combine with each other to form complex conjunct glyphs. Although the encoding encapsulates only the basic alphabetic characters, the number of glyphs and their combinations required for the exhaustive rendering of these scripts can be quite large (2000 - 4000+ glyph combinations).[2]

There are several contenders in the market giving unique solutions to representation of Indian languages but the users of any one of these cannot even exchange a file as they are proprietary font designs. It means the users have to acquire as many font libraries as there are vendors to be able to read the content. For web applications the problem is tackled by providing dynamic fonts but even this has the disadvantage of burdening networks with extra traffic and transmission costs.

**10.6.4.1 ISCI – Indian Script Code for Information Interchange**

The ISCI code standard specifies a 7-bit code table, which can be used in a 7 or 8-bit ISO compatible environment. It allows English and Indian script alphabets to be used simultaneously. It retains the ASCII character set in the lower half (0-127) of the 8-bit code table and provides Indian script characters in the upper half (160-255). ISCI caters to the following 10 Indian scripts - Devanagari, Gujarati, Punjabi, Bengali, Assamese, Oriya, Telugu, Tamil, Malayalam, Kannada. The ISCI code table is a superset of all the characters required for the above-mentioned scripts. First version was released in 1983 and adopted by the Bureau of Indian Standards (BSI) in 1991 after revisions in 1986 and 1988 [5].

#### **10.6.4.2 The GIST Card**

Graphics and Intelligence based Scripting Technology (GIST) is developed by CDAC, India. It supports major Indic languages. GIST technology uses ASCII code for character representation using the extended ASCII table [6]. GIST product uses underlying standards like ISCII (Indian Script code for Information Interchange), ISFOC (for font representation on screen) and INSCRIPT (common keyboard layout for Indian Script). GIST is developed on 8-bit encoding system i.e. it is an extended ASCII system. It uses same code for same vowel and consonant in different Indic languages (phonetic arrangement). Thus for transliteration it becomes very handy. It is required only to change the mode (or switches) of language. Software like iLeap work on the same principle and enable transliteration.

#### ***Web applications and GIST:***

GIST has enabled development of web-based solutions in Indian Languages. The software utility called ‘iPlugin’ facilitates viewing websites in Indian language through browsers. It provides facility to create chat rooms and email in any Indian language. The biggest advantage with iPlugin is that now need not to download the language font on the local machine. Necessary plugins directly get loaded on the machine when website is accessed, which is through ‘Dynamic fonts’.

#### **10.6.4.3 Constraints with ISCII and GIST card**

The major problem with ISCII is its character encoding. It uses 8-bit character. The codes can only be used in India. Even to convert from English to any Indic language it requires lot of labor. And from other language to any Indic language is very difficult. The reason for this is that ISCII uses the namespace between 128 and 255 to load the characters of any one Indian language. With this at best bi-lingual representation is possible, that too one language English and other an Indian Language. Truly trans-lingual representations are not possible where more than two Indian languages are needed. The GIST Card has the constraints of it being a hardware component. There are compatibility and upgrade constraints. It uses a physical switch, to load different languages because of its dependence on ISCII character set.

#### **Self Check Exercises**

**6) Describe the application of UNICODE in libraries.**

**7) What is ISCII?**

**8) What is GIST?**

Note: i. Write your answers in the space given below.

ii. Check your answers with the answer given at the end of this Unit.

---

---

---

---

---

---



- For future expansion surrogate area has been defined which can generate combinations of two 2-bytes packets.
  - UNICODE bothers about characters not glyphs.
  - It renders logical ordering for writing i.e. from left to right and in case of Hebrew, Urdu etc. right to left.
  - Combining characters also follow logical order. Combining characters are non-spacing characters, which form combination to represent a new character.
- 4) The features of UNICODE are:
- Sixteen bit character code
  - Efficiency
  - Characters not glyphs
  - Plain text
  - Logical order
  - Unification
  - Compatibility characters
  - Dynamic composition
  - Equivalent sequence
  - Convertibility
- 5) Compatibility characters are the characters, which actually would not have been encoded because these are the variant form of some characters, which are already encoded.
- 6) Libraries particularly Digital libraries can record information in the language and script of the source document. Even the bibliographic information can be represented in the language of the document. Once UNICODE is adopted, one can exploit the advantages of transliteration so that data can be transliterated wherever and whenever it is required.
- 7) ISCII stands for “Indian Script Code for Information Interchange”. It specifies a 7-bit code table, which can be used in a 7 or 8-bit ISO compatible environment. It allows English and Indian script alphabets to be used simultaneously.
- 8) GIST stands for “Graphics and Intelligence based Scripting Technology”. It is a 8-bit code which supports major Indic languages.

---

## 10.9 GLOSSARY OF KEY TERMS

---

<b>Glyph:</b>	Glyphs are images. It represents the shapes that characters can have when they are displayed.
<b>ASCII:</b>	American Standard Code for Information Interchange
<b>ISCI:</b>	Indian Script Code for Information Interchange
<b>GIST:</b>	Graphics and Intelligence based Scripting Technology
<b>INSCRIPT:</b>	Common keyboard layout for Indian Script

---

## 10.10 REFERENCES AND FURTHER READINGS

---

1. Oard, Douglas W. 1997. Serving Users in Many Languages: Cross-Language Information Retrieval for Digital Libraries. D-Lib Magazine, December, 1997.  
<http://www.dlib.org/dlib/december97/oard/12oard.html>
2. Internet scenario in India. <http://www.mithi.com/techback/main.html>
3. The UNICODE standard, Version 3.0, UNICODE consortium. Addison-Wesley, Massachusetts, 2000.
4. Tripathi, Aditya. Design and development of multilingual information retrieval system with numeric MARC. Thesis submitted to the University of Pune, Dept. of Library and Information Science, Pune. (Unpublished).
5. Technology developed for Indian languages.  
<http://www.tdil.gov.in/standards.htm#iscii>
6. CDAC - India on Multilingual Technology.  
<http://www.cdacindia.com/html/milingual.asp>
7. Ling Cao, etal. Searching heterogeneous multilingual bibliographic sources  
<http://www7.scu.edu.au/programme/posters/1874/com1874.htm>
8. Development of script of India.  
<http://www.geocities.com/Athens/Parthenon/2104/scripts.html>

---

## **Unit - 11 Human Computer Interfaces**

---

### **Structure**

- 11.0 Objectives
- 11.1 Introduction
- 11.2 Technology for Human Computer Interaction (HCI)
- 11.3 Ergonomics
- 11.4 User System Interface
- 11.5 W3C Guidelines
- 11.6 Digital Library Interfaces
  - 11.6.1 Usability of Digital Libraries
  - 11.6.2 Interfaces to Digital Library system
- 11.7 Visualization
- 11.8 Think map
- 11.9 Personalization
- 11.10 Summary
- 11.11 Answers to Self Check Exercises
- 11.12 Glossary of Key Terms
- 11.13 References and Further Readings

---

### **11.0 OBJECTIVES**

---

This lesson aims at explaining the following:

- Human Computer Interaction (HCI)
- Technology advances for HCI
- Ergonomics
- Technological development to facilitate HCI
- W3C guidelines for Web Content
- Digital Library Interfaces
- Explain Research Issues – Personalization and Visualization

---

### **11.1 INTRODUCTION**

---

Computers have found applications in practically every function from mundane routines to quite complex intellectual tasks. The main attraction is to make use of the speed with which processes and analysis can be accomplished by computers. However, we observe that different users interact differently with computers and expect certain features that help their transactions. If a system is complicated and expects the users to guess the how to do things surely such a system will not be popular. The best example is working with operating system (OS) at the command prompt. It involved learning all the commands available to be able to interact with the computers and also the exact syntax for each command. However, when windows environment was incorporated with OS it provided a Graphical User Interface (GUI) mode

where the users could issue commands by clicking on the mouse buttons. This made it much popular, as users did not have to memorize and issue commands. The same is true of Internet. One of biggest factors that contributed to popularity of the Internet is the introduction of the World Wide Web (WWW), which is the GUI interface to Internet. The capability to hyperlink web resources and make them available in a clickable browser environment has made it very popular and widespread.

The area of Human Computer Interaction (HCI) encompasses the study of what are types of users interactions with computers and the expectations of users of how a system should allow for different levels and ways of interaction.

The basic questions in designing HCI are [1]

- How users as individuals can use the systems designed for their use
- How this affects the work situations - jobs
- How organizational systems and the technology in an organization changes

---

## 11.2 TECHNOLOGY FOR HCI

---

Vannevar Bush, introduced the “memex,” a computing device using something like hyperlink technology to bring information to every user’s fingertips. Following the idea of memex, Douglas Engelbart invented the first “mouse,” which he called an “X-Y Position Indicator,” It was a little gizmo housed in a wooden box on wheels that moved around the desktop and took the cursor with it on the display. Engelbart described the mouse as being an integral part of a “graphical windowed interface,” and invented what he called "a windowed GUI".

"Graphical User Interface." A GUI is what computer types call the system of icons, taskbars, and other objects that our computers use to display and access information. [2]

The idea of direct manipulation of objects on a screen is integral to the concept of a graphic interface. In 1963 Ivan Sutherland, published the concept of “Sketchpad,” which directly manipulated objects on a CRT screen using a light pen. This was the first GUI (Graphical User Interface) long before the term was coined. [3]

In 1968 Engelbart created NLS (oNLine System), a hypermedia groupware system that used the mouse, the windowed GUI, hypermedia with object addressing and linking, and even an early version of video teleconferencing to some audience at Stanford University.

In the early 70s, Alan Kay and team developed an interactive object-oriented [4] programming language called Smalltalk. Smalltalk featured a graphical user interface (GUI) that looked similar to later iterations from both Apple and Microsoft.

The first real-life, usable GUI appeared in Xerox’s Alto computer, which was introduced in 1974 and was envisioned as a smaller, much more portable replacement for the mainframes of the time. The Alto, featured graphically driven applications. The Alto featured a bit-mapping display, which was essential for displaying graphics and WYSIWYG printing.

In 1981, the design and concepts which gave birth to the Alto led to the development and production of the much more streamlined, and more usable Xerox Star – the first true GUI-driven PC.

In 1984 Apple launched the Mac (MacIntosh) also bundled MacPaint, which brought computer “art” design to the average user and MacWrite, a simple word processor that was the first WYSIWYG (What You See Is What You Get) product of its kind on the consumer market. The Mac’s user-friendly interface made it popular at all levels of the computing community. In 1986, the desktop publishing application, PageMaker, for the Mac was released which was widely deployed for graphic arts and desktop publishing.

In 1985 Commodore launched its Amiga line of home PCs. The Amiga was the first PC to truly introduce the idea of “multimedia”. Amiga’s advanced sound and video capabilities went along with its sophisticated GUI-driven OS.

Microsoft Windows 1.0 made its official debut in November 1985 and Windows 2.0 in 1987 and version 3.0 in 1990. Around mid-80s, the first UNIX GUI, X-windows appeared as well. Later on the several versions of windows followed and the integrated MS Windows is now referred to as Operating System.

The main trend in the above mentioned developments was that most of these technology products are based upon the lessons learnt from Human Computer Interaction. Windows offered all the OS functions and utilities in a graphic mode that were much easier to use than in the command prompt mode. Moreover the functions are all available as menu based choices. Menus are a type of dialogue in which a user selects one item out of a list of displayed alternatives, whether the selection is by pointing, by entry of an associated option code, or by activation of an adjacent function key. Added to this is the semantics added through icons. Icons are small graphical representations of items in the computer such as files and folders or programs. Icons are particularly designed to convey what programs they represent. For example having an icon with a prominent ‘W’ for the MS-Word program shows that that is the icon used to invoke word. Icons are also used in programming where component in the form of icons can be dragged into the program as required.

Multimedia capability is one more attractive option that increases Human Computer interaction and ensures comfortable interaction by patrons. A popular adage goes that a picture can convey what a thousand words cannot.

Added to the software tools are the physical tools that ensure better interactions. A prominent one is the indisputable mouse. With the click of the mouse programs can be executed, data saved and world resources surfed! Advanced tools include touch screens, test readers and on screen plotters.

All the above-mentioned advances made HCI much more user friendly as compared to using computer tools a decade or two ago. But in complex systems such as digital libraries a lot of work still has to be done to improve upon their usability.

---

## 11.3 ERGONOMICS

---

Ergonomics is an area that involves study of building a suitable physical environment to facilitate comfortable work tools and environs. Long exposure to computers and using its peripherals has induced a lot of work related disorders. Occupational hazards are a topic that has long been discussed and it is even more emphasized with computer related hazards. Computer Ergonomics is the study of human capabilities in relationship to the specific work demands of the computer user. The word ergonomics is derived from the Greek words "ergon", which means work and "nomoi", which means natural law. Computer ergonomics is important to anyone who uses the computer for work. Ergonomics is the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and applies theory, principles, data and methods to design in order to optimize human well being and overall system performance. Ergonomists contribute to the design and evaluation of tasks, jobs, products, environments and systems in order to make them compatible with the needs, abilities and limitations of people. Ergonomics is often discussed as a part of Human Computer Interaction. In short it involves the design of equipment to match the working styles of people. Accordingly it studies the user physical requirement and degree of comfort and tolerance and makes recommendations in user centric design of workstations and providing adequate air, light, acoustics and physical environment. It seeks to eliminate the particularly high prevalence of repetitive stress injuries and work related disorders in the modern day office.

Following ergonomic principles helps to reduce stress and eliminate many potential injuries and disorders associated with the overuse of muscles, bad posture, and repeated tasks. This is accomplished by designing tasks, workspaces, controls, displays, tools, lighting, and equipment that fit the physical human capabilities and limitations [4]. Ergonomics promotes a holistic approach in which considerations of physical, cognitive, social, organizational, environmental and other relevant factors are taken into account.

### Self Check Exercise

#### 1) What is ergonomics? What is the purpose of it?

Note: i. Write your answers in the space given below.

ii. Check your answers with the answer given at the end of this Unit.

---

---

---

---

---

---

---

---

---

---

---

---

## 11.4 USER SYSTEM INTERFACE

---

What is the user-system interface? In common usage, the phrase is broadly defined to include all aspects of system design that affect system use (Smith, 1982a) [5].

In practical systems, User Interface (UI) is the module that allows users to use the system according to their need. Since UI is the only means for users to use the system it should incorporate the necessary navigational and assistance furnished in a way that users find it easy to use. There is no single solution to building a good UI as this component is based on human experience and the requirements may be varied and dynamic. Each user will have a unique path of interaction or even the same user may have a different path of action at every instance of using the system. Hence it is important to follow some guidelines while designing UIs so that they may cater to needs represented by a cross section of the user community.

To the extent that information systems support human users performing defined tasks, careful design of the user-system interface will be needed to ensure effective system operation. Users of information systems interact with a computer in order to accomplish information-handling tasks necessary to get their jobs done. They differ in ability, training and job experience. They may be keenly concerned with task performance, but may have little knowledge of (or interest in) the computers themselves. Design of the user-system interface must take account of those human factors. [5]

As mentioned earlier a system use depends upon on the efficiency of the user interface and interaction that are provided. Usability studies focus on factors that determine the extent of use of a particular system.

Usability can characterize any aspect of the ways that people interact with a system, even its installation and maintenance. The following attributes are given by Nielsen [15] to determine interface usability. He suggests that these attributes can be evaluated through usability testing relative to certain users and certain tasks. [6]

1. Learnability - Ease of learning such that the user can quickly begin using it.
2. Efficiency - Ability of user to use the system with high level of productivity.
3. Memorability - Capability of user to easily remember how to use the system after not using it for some period.
4. Errors - System should have low error rate with few user errors and easy recovery from them. Also no catastrophic errors.

### Self Check Exercise

**1) What is User Interface? What are the different attributes to determine interface usability?**

Note: i. Write your answers in the space given below.

ii. Check your answers with the answer given at the end of this Unit.



not provide sufficient contrast when viewed using monochrome displays or by people with different types of color deficits.

***Guideline 3. Use markup and style sheets and do so properly.***

Mark up documents with the proper structural elements. Using markup improperly and not adhering to specifications hinders accessibility. Misusing markup for a presentation effect (e.g., using a table for layout or a header to change the font size) makes it difficult for users with specialized software to understand the organization of the page or to navigate through it. Furthermore, using presentation markup rather than structural markup to convey structure (e.g., constructing what looks like a table of data with an HTML PRE element) makes it difficult to render a page intelligibly to other devices. Content developers may be tempted to use (or misuse) constructs that achieve a desired formatting effect on older browsers. They must be aware that these practices cause accessibility problems and must consider whether the formatting effect is so critical as to warrant making the document inaccessible to some users.

At the other extreme, content developers must not sacrifice appropriate markup because a certain browser or assistive technology does not process it correctly. For example, it is appropriate to use the TABLE element in HTML to mark up tabular information even though some older screen readers may not handle side-by-side text correctly. Using TABLE correctly and creating tables that transform gracefully makes it possible for software to render tables other than as two-dimensional grids.

***Guideline 4. Clarify natural language usage.***

Use markup that facilitates pronunciation or interpretation of abbreviated or foreign text.

When content developers mark up natural language changes in a document, speech synthesizers and Braille devices can automatically switch to the new language, making the document more accessible to multilingual users. Content developers should identify the predominant natural language of a document's content (through markup or HTTP headers). Content developers should also provide expansions of abbreviations and acronyms. In addition to helping assistive technologies, natural language markup allows search engines to find key words and identify documents in a desired language. Natural language markup also improves readability of the Web for all people, including those with learning disabilities and cognitive and physical disabilities.

***Guideline 5. Create tables that transform gracefully***

Ensure that tables have necessary markup to be transformed by accessible browsers and other user agents. Tables should be used to mark up truly tabular information ("data tables"). Content developers should avoid using them to layout pages ("layout tables"). Tables for any use also present special problems to users of screen readers. Some user agents allow users to navigate among table cells and access header and other table cell information. Unless marked-up properly, these tables will not provide user agents with the appropriate information.

***Guideline 6. Ensure that pages featuring new technologies transform gracefully.***

Ensure that pages are accessible even when newer technologies are not supported or are turned off. Although content developers are encouraged to use new technologies that solve problems raised by existing technologies, they should know how to make their pages still work with

older browsers and people who choose to turn off features. One example is the use of frames. When frames are used a note should also be given that the page should be viewed using which versions of available browsers.

***Guideline 7. Ensure user control of time-sensitive content changes.***

Some people with cognitive or visual disabilities are unable to read moving text quickly enough or at all. Movement can also cause such a distraction that the rest of the page becomes unreadable. Screen readers are unable to read moving text. When using utilities like web casters, which use push technology to send information to the screens they offer many features like giving running text (tickers) and various other animations like screen bursts to present information. But these features may distract users in their routine work. In the design of web pages care should be taken to use animations and moving text so that either users could control the movements or the speed should be set such that it is convenient for the users.

***Guideline 8. Ensure direct accessibility of embedded user interfaces.***

Ensure that the user interface follows principles of accessible design: device-independent access to functionality, keyboard operability, etc. When an embedded object has its "own interface", the interface -- like the interface to the browser itself -- must be accessible. If the interface of the embedded object cannot be made accessible, an alternative accessible solution must be provided.

***Guideline 9. Design for device-independence.***

Use features that enable activation of page elements via a variety of input devices. Device-independent access means that the user may interact with the user agent or document with a preferred input (or output) device -- mouse, keyboard, voice, etc. If, for example, a form control can only be activated with a mouse or other pointing device, someone who is using the page without sight, with voice input, or with a keyboard or who is using some other non-pointing input device will not be able to use the form. For example, providing text equivalents for image maps or images used as links makes it possible for users to interact with them without a pointing device.

***Guideline 10. Use interim solutions.***

Use interim accessibility solutions so that assistive technologies and older browsers will operate correctly. For example, older browsers do not allow users to navigate to empty edit boxes. Older screen readers read lists of consecutive links as one link. These active elements are therefore difficult or impossible to access. Also, changing the current window or popping up new windows can be very disorienting to users who cannot see that this has happened.

***Guideline 11. Use W3C technologies and guidelines.***

Use W3C technologies (according to specification) and follow accessibility guidelines. Where it is not possible to use a W3C technology, or doing so results in material that does not transform gracefully, provide an alternative version of the content that is accessible.

***Guideline 12. Provide context and orientation information.***

Provide context and orientation information to help users understand complex pages or elements. Grouping elements and providing contextual information about the relationships between elements can be useful for all users. Complex relationships between parts of a page



short are expected to give all services in the online environment without human interference and embody 'librarianship online'.

### **11.6.1 usability of digital libraries**

Digital libraries are potentially powerful tools, but their potential will be realized only if users are able to find the utilities that they need easily and will be able to use them easily. Existing libraries provide the core essential functionality as they serve structured repositories of multimedia documents, and allow documents to be added and retrieved from the collections. However quality of the interaction depends on such systems also satisfying various non-functional requirements that relate to usability. Usability studies of Digital Libraries attempt to study how to make digital libraries usable. In the words of Buchanan and Blandford, in the context of Human-Computer Interaction (HCI), the term "usable" can mean a range of things, including the following: [8]

- How efficiently and effectively users can achieve their goals with a system (for which it may be possible to apply performance measures)
- How easily users can learn to use the system ("learnability")
- How well the system helps the user avoid making errors, or recover from errors
- How much users enjoy working with the system — the quality of the user experience — or whether they find it frustrating and
- How well the system fits within the context in which it is used.

Digital libraries have different and distinctive types of users. They are classified by their level of interaction with the DLs as follows:

***End-users or Digital Library Patrons:*** these are the users of digital libraries just like the member of a traditional library. They are able to access a DL, browse or search and access resource based on the permissions and authentication required of that DL.

***Digital library Managers and Personnel:*** these are persons responsible for making available the Digital library collections and services. They require to interact with the system for functions like system installation and maintenance, updating, back-ups, printing and updations. The functions may also include reviewing process for the resources submitted to collections.

***Digital library Contributors:*** Digital Library content is not only input by a library but built more as online collaborations. Authors can interact with the DL system and submit resources online and keep track of its publishing in the repository. They also interact with the review process online and make corrections required and re-submit. DL resources are described using a standard for metadata. However, author may not know the nuances of the metadata standard and how exactly metadata is to be supplied. But metadata is essential for resource discovery in digital libraries. Hence, it essential for the system to have a component that assists the authors in providing correct metadata.

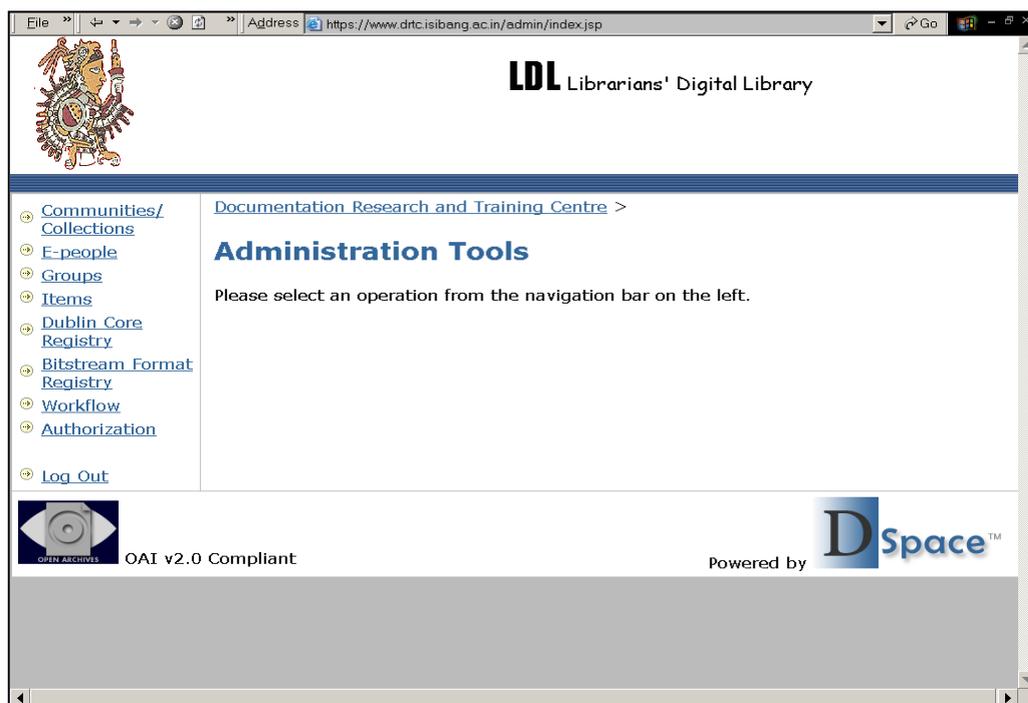
Based on the above discussion it becomes clear that interfaces to Digital libraries are complex and layered. DLs should make available the pertinent interface to the users depending upon what type of users one is and what type transactions he/she does with the system. This area is called designing of Digital Library Interfaces.

## 11.6.2 Interfaces to digital library system

As explained in the section above, there are different kinds of users of a digital library. Hence they should be provided with customizable interfaces to pertinent functions.

### Librarian Interface

This interface helps the Digital Librarians build and manage collections, manage users and user groups, manage the rights and permission to resources, updations and such administrative functions. Given below is illustration of the administrative tools available in the DSpace Digital Library Software. In DSpace the administrator is given browser based interface to make communities and collections, add users, authorized authors to submit resources, manage metadata elements for a collection and establish authorizations to collections. The advantage with such an interface is that it is browser based and can be managed from any machine connected to the server on which the actual DSpace system is loaded. The administrator can even delegate administrative work to others.

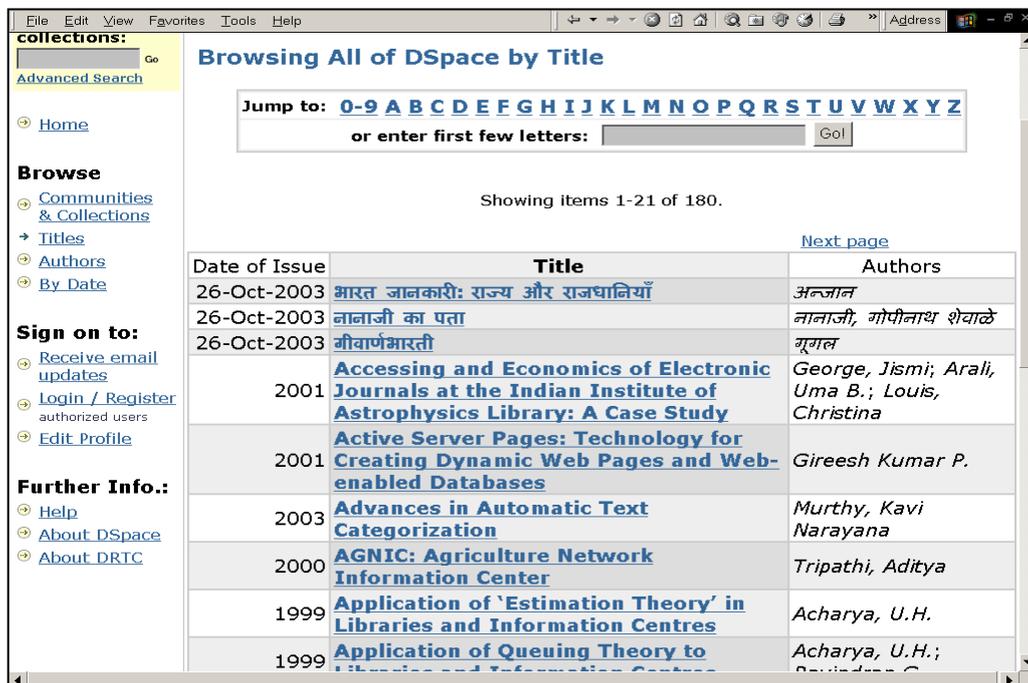


### Browse/Search Interface

End users of digital libraries need to browse and search the collections and retrieve resources as per their requirement. Digital libraries work in a distributed environment and often users have no human mediation to explain the contents or services as in the case of traditional libraries. Digital libraries imply librarianship online. All the tools that were used by human intermediaries must be made available as options to the end users in way that he/she can discover those and in a way that is easily usable. This is especially true of the browsing and the search interfaces. All the options desired should be made available with navigational facility in all directions. Figure below is a screenshot of the Browsing interface of the DSpace DL software. Collections can be browsed by communities and collection names, titles, authors or by date. A horizontal alphabetic bar is also provided to give alphabetic access to resources.

Once a user is in a collection, again it can be browsed by titles, author or date. This interface is also browser based.

The search interface allows searching by any term across collections. The advanced search facility allows searching by the metadata elements in Boolean combinations or using proximity operators or by fuzzy logic. The search capability is very powerful but the operations can be exercised by users using the interface without having to learn too much about how to execute what type of search.



## Metadata Interface

Authors submitting resources to Digital libraries are expected to furnish the metadata that best describes the content and helps in resource discovery. The standards that are followed for such descriptions are known as 'Metadata Standards'. Dublin Core is one popular standard being used to describe (mostly text based) resources in Digital Libraries. However, librarians are used to cataloguing standards and the art of making catalogue records that help users find resources by their approach terms. The same is not to be expected of the end users or the authors who wish to submit to Digital Libraries. Hence metadata should be as simple as possible and the terminology should be as generically defined as possible so that it is understood and can be interpreted across communities. However, it is necessary to provide an interface with help messages to author about how and what data to furnish for which element. Given below is the metadata form provided in DSpace for authors to furnish metadata is Dublin Core. Simple instruction about how to render an author's name can be provided. Wherever options can be made available for a particular element, a drop down list is provided so that authors can pick up the relevant ones. DL manager can add to or delete items from such lists.

Submit: Describe Your Item

Please fill in the requested information about your submission below. In most browsers, you can use the tab key to move the cursor to the next input box or button, to save you having to use the mouse each time. ([More Help...](#))

Enter the names of the authors of this item below.  
*Last name*                      *First name(s) + "Jr"*  
*e.g. Smith*                      *e.g. Donald Jr*

**South Indian & Similar:**  
*First name*                      *Initials*  
*e.g. Madhusudan*                      *e.g. C.*

**Authors**

**Title**

Enter the series and number assigned to this item by your community.  
*Series Name*                      *Report or Paper No.*

**Series/Report No.**

If the item has any identification numbers or codes associated with it, please enter the types and the actual numbers or codes below.

**Identifiers**

Select the type(s) of content you are submitting. To select more than one value in the list, you may have to hold down the "CTRL" or "Shift" key.

**Type**

**Self Check Exercise**

**4) Who are the different users of a Digital Library?**

Note: i. Write your answer in the space given below.

ii. Check your answer with the answer given at the end of this Unit.

---



---



---



---



---

**11.7 VISUALIZATION**

Visualization is an area dealing with visualizing user interactions and design according to human perception of system. It also refers to providing visual tools for information presentation and access. In the words of Hearst [9], Information visualization is a complex research area. It builds on theory in information design, computer graphics, human-computer interaction and cognitive science. Practical application of information visualization in computer programs involves selecting, transforming and representing abstract data in a form that facilitates human interaction for exploration and understanding. Important aspects of information visualization are the interactivity and dynamics of the visual representation. Strong techniques enable the user to modify the visualization in real-time, thus affording unparalleled perception of patterns and structural relations in the abstract data in question. [9]

An example of incorporating visualization tools is that of Grokker. Grokker displays colored circles within circles (or squares within squares), grouping classes and subclasses. Mouse-over display of the metadata accompanies each item such as print books, web sites, ebooks, and videos. Users can sort results by domain and customize and store results for future use. Flexible filters that adjust to the source being searched make it easy to narrow a search on the fly.

Another area of visualizations pertains to visualizing information representation and display of the search results of a query. Some graphical tools are available that graphical map concepts into nodes and at each node other concepts are linked according to their relationships with other concepts. Thus it is essentially a visual knowledge map that can be accessed at each node.

---

## **11.8 THINKMAP**

---

Thinkmap provides solutions for displaying, animating, and navigating complex and interconnected information. [10] Thinkmaps are custom interfaces that transform data and information into insight and knowledge. Thinkmap interfaces present data in a way that makes visible both data and the interconnections between data.

---

## **11.9 PERSONALIZATION**

---

Libraries have long strived to offer tailored services to their patrons. In the parlance of Digital Libraries (DLs) there is the added challenge of offering customized services to users in the digital online world. The research area dealing with viable solutions to such customized interfaces to DLs is known as 'Personalization'. Personalization is the way in which information and services can be tailored to match the unique and specific needs of an individual or a community. This is achieved by adapting presentation, content, and/or services based on a person's task, background, history, device, information needs, location, etc., essentially the user's context.

One of the products of personalized services is a Recommender system. Recommender systems are a particular type of personalization that learn about a person's needs and then proactively identify and recommend information that meets those needs. Recommender systems are especially useful when they identify information a person was previously unaware of. Personalization can be user-driven which involves a user directly invoking and supporting the personalization process by providing explicit input. Some products have been deployed to give personalized services. One commonly available example is MyYahoo! where the user explicitly initiates actions and provides example information in order to control the personalization. [11] PIE (Personalized Information Environment) (Jayawardana 2001) in a digital library is a framework that provides a set of integrated tools based on an individual user's requirements and interests with respect to his access to library materials. These tools can support active learning by integrating the user's personal library and a remote digital library. The user will be able to carry out learning activities when browsing the digital library. [12]

## Self Check Exercise

5) What is personalization?

6) What is Recommender system?

Note: i. Write your answers in the space given below.

ii. Check your answers with the answer given at the end of this Unit.

---

---

---

---

---

---

---

---

---

---

## 11.10 Summary

A system is only as good as its usage by its users. The best content may be put into digital libraries but unless very good interfaces are made available the content will not be accessed. It is not just the lay users but practically most users get discouraged to use a system that pre-supposes technical knowledge of the users. The design should cater to the lowest common denominator of understanding and usage of computerized systems. Digital libraries are complex having several modules intended for different kinds of users. At every stage it should envisaged what types of users may interact with it and pertinent interface should be made available. Users should be intuitively led from one logical step to the next. All such factors that aid Human Computer Interaction should be incorporated into design of Digital libraries. User-friendly systems are accomplished in user centric design of digital library software and systems, rather than machine centric design.

---

## 11.11 ANSWERS TO SELF CHECK EXERCISES

---

1) Ergonomics is the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and applies theory, principles, data and methods to design in order to optimize human well being and overall system performance.

It helps to reduce stress and eliminate many potential injuries and disorders associated with the overuse of muscles, bad posture, and repeated tasks.

2) User Interface (UI) is the module that allows users to use the system according to their need.

The attributes are, Learnability, Efficiency, Memorability, Errors

- 3) The guidelines for presentation of information on the Web are,  
Guideline 1: Provide equivalent alternatives to auditory and visual content  
Guideline 2. Don't rely on color alone  
Guideline 3. Use markup and style sheets and do so properly  
Guideline 4. Clarify natural language usage  
Guideline 5. Create tables that transform gracefully  
Guideline 6. Ensure that pages featuring new technologies transform gracefully  
Guideline 7. Ensure user control of time-sensitive content changes  
Guideline 8. Ensure direct accessibility of embedded user interfaces  
Guideline 9. Design for device-independence  
Guideline 10. Use interim solutions  
Guideline 11. Use W3C technologies and guidelines  
Guideline 12. Provide context and orientation information  
Guideline 13. Provide clear navigation mechanisms  
Guideline 14. Ensure that documents are clear and simple
- 4) The users of a Digital Library are of three types,
  - a) End-users or Digital Library Patrons
  - b) Digital Library Managers and Personnel
  - c) Digital Library Contributors
- 5) Personalization is the way in which information and services can be tailored to match the unique and specific needs of an individual or a community.
- 6) Recommender systems are a particular type of personalization that learn about a person's needs and then proactively identify and recommend information that meets those needs.

---

## 11.12 GLOSSARY OF KEY TERMS

---

<b>HCI</b>	Human Computer Interaction
<b>GUI</b>	Graphical User Interface
<b>WWW</b>	World Wide Web
<b>oNLine System (NLS)</b>	A hypermedia groupware
<b>User Interface (UI)</b>	The module that allows users to use the system according to their needs.
<b>W3C</b>	World Wide Web Consortium
<b>HTTP</b>	HyperText Transfer Protocol. It is the primary method used to convey information on the World Wide Web.
<b>DSpace</b>	It is an Open Source Digital Library Software.

**Dublin Core (DC)**

A metadata standard

**PIE**

Personalized Information Environment

---

## 11.13 REFERENCES AND FURTHER READINGS

---

1. Components of HCI, at <http://www.uwasa.fi/~mj/hci/hci3.html>
2. SitePoint Glossary, at <http://www.sitepoint.com/glossary.php?q=G>
3. Biography of a Luminary Dr. Ivan E. Sutherland, at [http://www.cc.gatech.edu/classes/cs6751\\_97\\_fall/projects/abowd\\_team/ivan/ivan.html](http://www.cc.gatech.edu/classes/cs6751_97_fall/projects/abowd_team/ivan/ivan.html)
4. What is ergonomics? At <http://www.cdc.gov/od/ohs/Ergonomics/ergodef.htm>
5. Ref. 4
6. Nielsen, J. (1993). Usability Engineering. Boston, MA: Academic Press, Inc.
7. Web Content Accessibility Guidelines 1.0. W3C Recommendation 5-May-1999. <http://www.w3.org/TR/WAI-WEBCONTENT/>
8. Usability of digital libraries: a source of creative tensions with technical developments, at <http://www.ieee-tcdl.org/Bulletin/current/blandford/blandford.html>
9. Hearst, Marti. User Interfaces and Visualization, at <http://www.sims.berkeley.edu/~hearst/irbook/10/node5.html#fig:themescapes>
10. Thinkmap, at <http://www.thinkmap.com/>
11. Personalization tools for active learning in digital libraries, at <http://wings.buffalo.edu/publications/mcjrnl/v8n1/active.pdf>
12. Jayawardana, Champa; Hewagamage, K. Priyantha and Hirakawa, Masahito. Personalization tools for active learning in digital libraries. MC Journal: The Journal of Academic Media Librarianship, Vol.8, No.1, Summer 2001. at <http://wings.buffalo.edu/publications/mcjrnl/v8n1/active.pdf>
13. Smith, S. L., and Mosier, J. N. (1984a). Design Guidelines for User-System Interface Software (Technical Report ESD-TR-84-190). Hanscom Air Force Base, MA: USAF Electronic Systems Division. (NTIS No. AD A154 907) as quoted in <http://hcibib.org/sam/> retrieved on 25/10/2005)
14. Wikipedia, at [http://en.wikipedia.org/wiki/Information\\_visualization](http://en.wikipedia.org/wiki/Information_visualization)

---

## **Unit – 12 : Resource Discovery**

---

### **Structure**

- 12.0 Objectives
- 12.1 Introduction
- 12.2 Internet and Resource Discovery
- 12.3 Standards for Resource Discovery
- 12.4 Metadata
- 12.5 Dublin Core Standard
- 12.6 Role of XML (eXtensible Markup Language) in Resource Discovery
- 12.7 Resource Discovery Framework (RDF)
- 12.8 Summary
- 12.9 Answers to Self Check Exercises
- 12.10 Glossary of Key Terms
- 12.11 References and Further Readings

---

### **12.0 OBJECTIVES**

---

Reading this lesson will help to understand:

- Metadata
- Dublin Core Standard
- Resource Discovery Framework
- Role of XML in Resource Discovery

---

### **12.1 INTRODUCTION**

---

The term ‘Resource Discovery’ refers to locating and retrieving resources from a collection. The task is not new to librarianship. A library organizes collection systematically and provides tools that help patrons access particular items in the collection by different approaches. A traditional library usually provides a systematic approach through tools such as the catalogue. A catalogue is a collection of bibliographic records of the library holdings and allows the users to search by author, title, publisher and most popular subject approach.

Now consider online electronic collections. The question arises whether the same kind of traditional organization and retrieval tools are available for locating items in the large collection of resources.

---

### **12.2 INTERNET AND RESOURCE DISCOVERY**

---

The general search engines on Internet often return millions of hits of a given query, which is like retrieving the entire net for a query. The level of satisfaction is very less

with lot of recall and literally no precision in web retrieval. This leads to research in implementing organization tools and techniques to aid efficient retrieval instead of too much noise or unwanted hits. The objective is to build meaningful search techniques to aid resource discovery. Resource discovery is the term commonly used to refer to the exercise of locating, accessing, retrieving, and managing relevant resources from widely distributed heterogeneous networks [1].

### **Self Check Exercises**

#### **1) What is resource discovery?**

Note: i. Write your answers in the space given below.

ii. Check your answers with the answer given at the end of this Unit.

---

---

---

---

---

---

---

---

---

### **12.3 STANDARDS FOR RESOURCE DISCOVERY**

---

The main reason for poor retrieval on the Internet is the unorganized manner in which information is presented. HTML, the language in which webpages are written, mostly deals with structural information and not description of the content of the tags. Hence there is no clue for the search, where or what part of document means what. With the advent of XML a new paradigm is presented where meaningful tag sets with domain specific vocabularies can be created. This has brought quite some excitement in communities that are putting forward their own elements sets as per their requirement. Again this causes a lot variation in how web/electronic resources are represented and does not help achieve resource discovery. Hence metadata standards have been proposed and adapted as standards that help in uniformly describing resources in order to ultimately achieve precision in resource discovery.

---

### **12.4 METADATA**

---

Metadata is data about data just like cataloguing data or bibliographic records. The effectiveness of resource discovery systems will rely on flexible and extensible naming and metadata mechanisms as a key to accessing resources. There is a need for describing the Internet or web based documents, so that search and retrieval are effective and efficient. A simple standard for cross-domain resource discovery is the Dublin Core Standard.

#### **Dublin Core**

The Dublin Core Metadata Initiative (DCMI) is an organization dedicated to fostering the widespread adoption of interoperable metadata standards and promoting the

development of specialized metadata vocabularies for describing resources to enable intelligent resource discovery systems. Dublin Core metadata provides card catalog-like definitions for defining the properties of objects for Web-based resource discovery systems.

## Dublin Core and HTML

Web pages are one of the most common types of resources to utilize the Dublin Core's descriptions, usually within HTML's meta tags. There are many digital archives of physical objects that also use Dublin Core. Dublin Core metadata is often stored as name-value pairs within META tags, which are placed within the HEAD element of an HTML document. However, it can also be located in an external document or loaded into a database enabling it to be indexed and manipulated from within a propriety application. But the latest trend is to use XML tags rather than HTML. In the section 11.5.1, examples are given for embedding DC elements in HTML and also in XML tags.[2]

### Meta Tag

The META tag of HTML is designed to encode a named metadata element. Each element describes a given aspect of a document or other information resource. For example, this tagged metadata element,

```
<meta name = "DC.Creator"  
      content = "Simpson, Homer">
```

says that Homer Simpson is the Creator, where the element named Creator is defined in the DC element set. In the more general form,

```
<meta name = "PREFIX.ELEMENT_NAME"  
      content = "ELEMENT_VALUE">
```

the capitalized words are meant to be replaced in actual descriptions; thus in the example,

```
ELEMENT_NAME was: Creator  
ELEMENT_VALUE was: Simpson, Homer  
and PREFIX was: DC
```

Within a META tag the first letter of a Dublin Core element name is capitalized. DC places no restriction on alphabetic case in an element value and any number of META tagged elements may appear together, in any order. More than one DC element with the same name may appear, and each DC element is optional.

---

## 12.5 DUBLIN CORE STANDARD

---

Dublin Core consists of the following:

- **Elements**

An Element is a property of a resource. E.g. Title, author, publisher, etc are properties of a document.

- **Qualifiers**

Qualifiers are used to narrow the scope of an element. DC elements can be used without qualifiers. Qualifiers are of two types as below:

- **Element Refinements**

Element refinements are used to sharpen focus of a concept by further qualifying it. Classic subject indexing example is ROSE qualified by color 'red' in term 'red rose'.

- **Encoding Schemes**

Encoding schemes provide contextual information or parsing rules that aid in the interpretation of a term value. For example, controlled vocabularies, formal notations, or parsing rules. Encoding schemes are again two types:

1. Vocabulary Encoding Schemes

Enables providing standard vocabulary from different sources such as LCSH, MESH, etc

2. Syntax Encoding Schemes

Encoding schemes are used wherever the entities being described conform to a standard. For example for date the world standard format is given by W3C DTF (YYYY-MM-DD format)

- **DCMI Type**

The types of resources described are listed under the DCMI types include the following:

- Collection
- Dataset
- Event
- Image
- Interactive Resource
- Moving Image
- Physical Object
- Service
- Software
- Sound
- Still Image
- Text

### 12.5.1 Dublin core elements

The core elements of Dublin core are -- Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights, Provenance, Audience, rights Holder. These 18 elements are described below [2]:

#### *Element: Title*

Name: Title

Identifier: Title

Definition: A name given to the resource.

Comment: Typically, a Title will be a name by which the resource is formally known.

*Examples:*

#### **HTML:**

```
<meta name = "DC.Title"
      content = "Introduction to Internet">
```

#### **XML:**

```
<DC:Title>"Introduction to Internet"</DC:Title>
```

#### *Element: Creator*

Name: Creator

Identifier: Creator

Definition: An entity primarily responsible for making the content of the resource.

Comment: Examples of a Creator include a person, an organisation, or a service.

Typically, the name of a Creator should be used to indicate the entity.

*Examples:*

#### **HTML:**

```
<meta name = "DC.Creator"
      content = "Ranganathan, S.R.">
```

#### **XML:**

```
<DC:Creator>"Russel, Betrand"</DC:Creator>
```

#### *Element: Subject*

Name: Subject and Keywords

Identifier: Subject

Definition: The topic of the content of the resource.

Comment: Typically, a Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource.

Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.

*Examples:*

**HTML:**

```
<meta name = "DC.Subject"
      scheme = "MESH"
      content = "Carcenoma"> -- (this example shows how a vocabulary list can
be adopted for DC).
```

**XML:**

```
<DC:Subject>Library Classification</DC:Subject>
```

***Element: Description***

Name: Description

Identifier: Description

Definition: An account of the content of the resource.

Comment: Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.

*Examples:*

**HTML:**

```
<meta name = "DC.Description"
      lang = "en"
      content = "The author presents a brief overview of metadata with extensive
examples. ">
```

**XML:**

```
<DC:Description>The author presents a brief overview of metadata with extensive
examples. </DC:Description>
```

***Element: Publisher***

Name: Publisher

Identifier: Publisher

Definition: An entity responsible for making the resource available

Comment: Examples of a Publisher include a person, an organisation, or a service.

Typically, the name of a Publisher should be used to indicate the entity.

*Examples:*

**HTML:**

```
<meta name = "DC.Publisher"
      content = "Wrox">
```

**XML:**

```
<DC:Publisher>Wrox </DC:Publisher>
```

***Element: Contributor***

Name: Contributor

Identifier: Contributor

Definition: An entity responsible for making contributions to the content of the resource.

Comment: Examples of a Contributor include a person, an organization, or a service.

Typically, the name of a Contributor should be used to indicate the entity.

*Examples:*

**HTML:**

```
<meta name = "DC.Contributor.Artist"
      content = "Laxman, R.K.">
```

**XML:**

```
<DC:Contributor.refinement="Artist"> Laxman, R.K. </DC:Contributor>
```

***Element: Date***

Name: Date

Identifier: Date

Definition: A date associated with an event in the life cycle of the resource.

Comment: Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [[W3CDTF](#)] and follows the YYYY-MM-DD format.

*Examples*

**HTML:**

```
<meta name = "DC.Date"
      content = "1990">
```

```
<meta name = "DC.Date"
      content = "1990-05-14">
```

```
<meta name = "DC.Date.Created"
      content = "1990-05-14">
```

**XML:**

```
<DC:Date>1998-05-21<dc:date refinement="available">2002-06</dc:date>
```

***Element: Type***

Name: Resource Type

Identifier: Type

**Definition:** The nature or genre of the content of the resource.

**Comment:** Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the working draft list of Dublin Core Types [[DCT1](#)]). To describe the physical or digital manifestation of the resource, use the **FORMAT** element.

*Examples:*

**HTML:**

```
<meta name = "DC.Type"
      content = "drama">
```

```
<meta name = "DC.Type"
      content = "software">
```

**XML:**

```
<DC.Type> interactive video game</DC.Type>
```

**Element: Format**

Name: Format

Identifier: Format

**Definition:** The physical or digital manifestation of the resource.

**Comment:** Typically, Format may include the media-type or dimensions of the resource. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types [[MIME](#)] defining computer media formats).

*Examples:*

**HTML:**

```
<meta name = "DC.Format"
      content = "text/xml">
```

```
<meta name = "DC.Format"
      scheme = "IMT"
      content = "text/xml">
```

**XML:**

```
<DC.Format.>text/xml</DC:Format.scheme>
```

**Element: Identifier**

Name: Resource Identifier

Identifier: Identifier

Definition: An unambiguous reference to the resource within a given context.

Comment: Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system.

Example: Formal identification systems include the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).

*Examples:*

**HTML:**

```
<meta name = "DC.Identifier"  
  content = "http://www.dublincore.org/">
```

```
<meta name = "DC.Identifier"  
  scheme = "ISBN"  
  content = "1-56592-149-6">
```

**XML:**

```
<DC.Identifier>http://loc.gov.in/</DC.Identifier>  
<DC:Identifier> 1-56592-149-6</DC:Identifier>
```

***Element: Source***

Name: Source

Identifier: Source

Definition: A Reference to a resource from which the present resource is derived.

Comment: The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.

*Examples:*

**HTML:**

```
<meta name = "DC.Source"  
  content = "Bernard Shaw's Saint Joan">
```

**XML:**

```
<DC:Source> Bernard Shaw's Saint Joan </DC:Source>
```

***Element: Language***

Name: Language

Identifier: Language

Definition: A language of the intellectual content of the resource.

Comment: Recommended best practice for the values of the Language element is defined by RFC 1766 [RFC1766] which includes a two-letter Language Code (taken from the ISO 639

standard [ISO639]), followed optionally, by a two-letter Country Code (taken from the ISO 3166 standard [ISO3166]). For example, 'en' for English, 'fr' for French, or 'en-uk' for English used in the United Kingdom.

*Examples:*

**HTML:**

```
<meta name = "DC.Language"
      content = "en">
```

**XML:**

```
<DC:Language.scheme = "ISO639-2" >eng </DC:Language.scheme>
```

***Element: Relation***

Name: Relation

Identifier: Relation

Definition: A reference to a related resource.

Comment: Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.

*Examples:*

**HTML:**

```
<meta name = "DC.Relation.IsPartOf"
      content = "http://foo.bar.org/abc/proceedings/1998/">
```

**XML:**

```
<DC:Relation.> Shakespeare's Romeo and Juliet
      </DC:Relation.IsBasedOn>
```

***Element: Coverage***

Name: Coverage

Identifier: Coverage

Definition: The extent or scope of the content of the resource.

Comment: Coverage will typically include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity).

Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names [TGN]) and that, where appropriate, named places or time periods be used in preference to numeric identifiers such as sets of coordinates or date ranges.

*Examples:*

**HTML:**

```
<meta name = "DC.Coverage"
      content = "US civil war era; 1861-1865">
```

```
<meta name = "DC.Coverage"
```

content = "Columbus, Ohio, USA; Lat: 39 57 N Long: 082 59 W">

**XML:**

<DC:Coverage>Commonwealth Countries</DC:Coverage>

**Element: Rights**

Name: Rights Management

Identifier: Rights

Definition: Information about rights held in and over the resource.

Comment: Typically, a Rights element will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the Rights element is absent, no assumptions can be made about the status of these and other rights with respect to the resource.

Examples:

**HTML:**

```
<meta name = "DC.Rights"
  content = "Access limited to members">
```

**XML:**

<DC:Rights>Access limited to members</DC:Rights>

**Element: Audience**

Name: Audience

Identifier: Audience

*Definition:* A class of entity for whom the resource is intended or useful. A class of entity may be determined by the creator or the publisher or by a third party

*Comment:* Audience terms are best utilized in the context of formal or informal controlled vocabularies. None are presently recommended or registered by DCMI, but several communities of interest are engaged in setting up audience vocabularies. In the absence of recommended controlled vocabularies, implementers are encouraged to develop local lists of values, and to use them consistently.

*Examples:*

**HTML:**

```
<meta name = "DC.Audience"
  content = " elementary school students ">
```

**XML:**

<DC:Audience> elementary school students </DC:Audience>

### ***Element: Rights Holder***

Name: Rights Holder

Identifier: Rights Holder

Definition: A person or organization owning or managing rights over the resource. Recommended best practice is to use the URI or name of the Rights Holder to indicate the entity.

Comments: Since, for the most part, people and organizations are not typically assigned URIs, a person or organization holding rights over a resource would be named using a text string. People and organizations sometimes have websites, but URLs for these are not generally appropriate for use in this context, since they are not clearly identifying the person or organization, but rather the location of a website about them.

*Examples:*

#### **HTML:**

```
<meta name = "DC.RightsHolder"
      content = "Karnataka State Open University ">
```

#### **XML:**

```
<DC:RightsHolder> Karnataka State Open University </DC:RightsHolder>
```

### ***Element: Provenance***

Name: Provenance

Identifier: Provenance

Definition: A statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity and interpretation. The statement may include a description of any changes successive custodians made to the resource.

Comment: This especially important to clear the legal issues regarding the ownership and also authenticity of a resource and details like who makes it available in a repository, to who the resource belonged etc.

*Examples:*

#### **HTML:**

```
<meta name = "DC.Provenance"
      content = "This copy once owned by Benjamin Spock ">
```

#### **XML:**

```
<DC:Provenance>This copy once owned by Benjamin Spock</DC:Provenance>
```

### **Self Check Exercises**

**2) What is Qualifiers? Mention different types of qualifires.**

**3) What are the different elements of Dublin Core?**

Note: i. Write your answers in the space given below.

ii. Check your answers with the answer given at the end of this Unit.



**Definition:**

“A formal data model from the World Wide Web Consortium (W3C) for machine understandable metadata used to provide standard descriptions of web resources. It uses eXtensible Markup Language (XML). It is similar in intent to the Dublin Core, although perhaps broader in its scope and purpose.” [3].

The RDF model provides the description of Web documents (in other words rendering of metadata to the Web documents) in a neutral manner so that the metadata can be shared across different applications. It is essential to note that though XML is being used to represent documents in RDF, other languages also can emerge in future that can be used to describe documents in RDF.

**12.7.1 Features of RDF**

The most important feature of RDF is that it is developed to be domain-independent i.e. it is very general in nature and does not restrict/ apply any constraint on any one particular domain. It can be used to describe information about any domain. The RDF model imitates the class system of object-oriented programming. A collection of classes (as defined for a specific purpose or domain) is called a ‘*schema*’ in RDF. These classes are extensible through ‘*subclass refinement*’. [4] Thus, various related schemas can be made using the base schema. RDF also supports metadata reuse by allowing sharability between various schemas.

**12.7.2 Application areas [4]:**

The developers of RDF visualize RDF being used:

- in resource discovery to provide better search engine capabilities,
- in cataloging for describing the content and content relationships available at a particular Web site, page, or digital library,
- by intelligent software agents to facilitate knowledge sharing and exchange, in content rating,
- in describing collections of pages that represent a single logical "document",
- for describing intellectual property rights of Web pages, and
- for expressing the privacy preferences of a user as well as the privacy policies of a Web site. RDF with digital signatures will be key to building the "Web of Trust" for electronic commerce, collaboration, and other applications.

A simple RDF model has three parts [5]:

- i. Resource: Any entity, which has to be described, is known as Resource also known as *Subject*. It can be a ‘webpage’ on Internet or a ‘person’ in a society.
- ii. Property: Any characteristic of Resource or its attribute, which is used for the description of the same, is known as Property, also known as *Predicate*. For example, a web page can be recognized by ‘Title’ or a man can be

- recognized by his 'Name'. So both are attributes for recognition of resource 'webpage' and 'person' respectively.
- iii. Value: A Property must have a value also known as *Object*. Like, the title of DRTC webpage is 'Documentation Research and Training Centre', name of a person is 'Ranganathan'.

The resource, property and value together are known as a 'Statement'. The statement formed by the example given above can be diagrammatically represented as in Fig. 1.



Fig. 1

A resource can have an identifier i.e. URI (Uniform Resource Identifier). A URI can be a URL, example <http://drtc.isibang.ac.in>. A Property itself can be a 'resource', which makes a complex representation model of the data. For example, say Creator of the webpage <http://drtc.isibang.ac.in> is 'Biswanath' who is Research Fellow at DRTC and has email [biswanath@drtc.ac.in](mailto:biswanath@drtc.ac.in). It can be represented diagrammatically as in Fig.2.

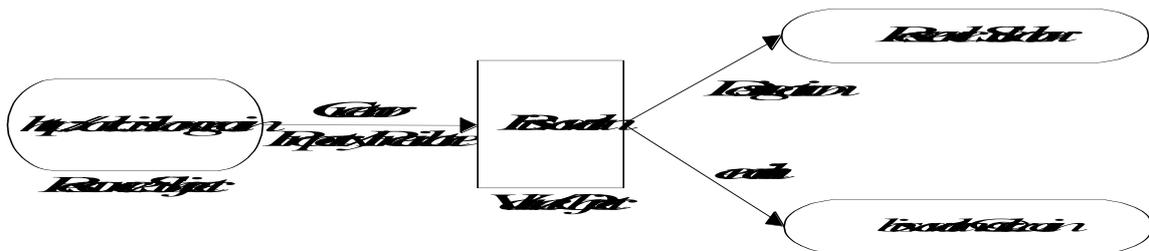


Fig. 2

### Self Check Exercises

- 4) What is RDF?
- 5) Discuss about the application areas of RDF.
- 6) Explain a simple RDF model.

Note: i. Write your answers in the space given below.

ii. Check your answers with the answer given at the end of this Unit.



- 2) Qualifiers are used to narrow the scope of an element. DC elements can be used without qualifiers.  
Qualifiers are of two types: Element Refinements and Encoding Schemes.
- 3) There are 18 Dublin Core elements. These are -- Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights, Provenance, Audience, rightsHolder.
- 4) A formal data model from the World Wide Web Consortium (W3C) for machine understandable metadata used to provide standard descriptions of web resources. It uses eXtensible Markup Language (XML). It is similar in intent to the Dublin Core, although perhaps broader in its scope and purpose.”
- 5) The developers of RDF visualize RDF being used:
  - in resource discovery to provide better search engine capabilities,
  - in cataloging for describing the content and content relationships available at a particular Web site, page, or digital library,
  - by intelligent software agents to facilitate knowledge sharing and exchange, in content rating,
  - in describing collections of pages that represent a single logical "document",
  - for describing intellectual property rights of Web pages, and for expressing the privacy preferences of a user as well as the privacy policies of a Web site. RDF with digital signatures will be key to building the "Web of Trust" for electronic commerce, collaboration, and other applications.
- 6) A simple RDF model has three parts:
  - Resource: Any entity, which has to be described, is known as Resource also known as *Subject*. It can be a ‘webpage’ on Internet or a ‘person’ in a society.
  - Property: Any characteristic of Resource or its attribute, which is used for the description of the same, is known as Property, also known as *Predicate*. For example, a web page can be recognized by ‘Title’ or a man can be recognized by his ‘Name’. So both are attributes for recognition of resource ‘webpage’ and ‘person’ respectively.
  - Value: A Property must have a value also known as *Object*. Like, the title of DRTC webpage is ‘Documentation Research and Training Centre’, name of a person is ‘Ranganathan’.

---

## 12.10 GLOSSARY OF KEY TERMS

---

**XML:** Extensible Markup Language

**Metadata:** Metadata is data about data just like cataloguing data or bibliographic records.

**DCMI:** Dublin Core Metadata Initiative.

**Dublin Core:** It is a metadata standard for describing digital objects (including webpages) to enhance visibility, accessibility and interoperability.

**Elements:** An Element is a property of a resource. e.g. Title, author, publisher, etc are properties of a document.

**RDF:** Resource Discovery Framework.

---

## 12.11 REFERENCES AND FURTHER READINGS

---

1. Wood, Andrew, Ward, Nigel, Sue, Hoylen, Iannella, Renato. Resource Discovery and the Open Information Locator Project.  
<http://www.w3.org/Search/9605-Indexing-orkshop/Papers/Wood@DSTC.html>
2. Powell, A and Johnston, A. Guidelines for implementing Dublin Core in XML. at  
<http://www.ukoln.ac.uk/metadata/dcmi/dc-xml-guidelines/>
3. Koch T. The role of classification schemes in Internet resource description and discovery. <http://www.ukoln.ac.uk/metadata/desire/classification/>
4. Glossary of Internet Terms. [www.walthowe.com/glossary/r.html](http://www.walthowe.com/glossary/r.html)
5. Prasad, A. R. D. and Patel, Dimple. An Overview of the Semantic Web.  
<http://drtc.isibang.ac.in/~dimple/semanticweb.pdf>
6. Resource Description Framework (RDF) Model and Syntax Specification W3C Recommendation 22 February 1999.  
<http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/#intro>
7. Dublin Core Official Site and usage guide.  
<http://dublincore.org/documents/usageguide/elements.shtml>
8. Dublin Core Metadata Initiative. <http://dublincore.org/>



**KARNATAKA STATE OPEN UNIVERSITY**  
MUKTHAGANGOTRI, MYSORE –570 006

**MASTER OF LIBRARY AND INFORMATION SCIENCE**  
**M.Lib.I.Sc - 5**

**Information Systems:  
Architecture and Retrieval**

**BLOCK - 4**

**BLOCK**

**4**

---

**MODELS OF INFORMATION RETRIEVAL & &  
EVALUATION OF RETRIEVAL SYSTEMS WITH  
SPECIAL REFERENCE TO ELECTRONIC LIBRARIES  
AND INFORMATION SYSTEMS.**

---

---

**Unit - 13**

**Philosophical approaches. Empirical Model/ Statistical Models/  
Probabilistic Models**

---

**Unit - 14**

**Cognitive Models and Application of Expert Systems in  
Information Retrieval: I<sup>3</sup> R, Cansearch, Plexus etc**

---

**Unit – 15**

**Retrieval effectiveness**

---

**Unit – 16**

**Retrieval efficiency & Evaluation studies**

## INSTRUCTIONAL DESIGN AND EDITORIAL COMMITTEE

### COURSE DESIGN

**Prof. D. Shivalingaiah**

**Chairman**

Vice Chancellor  
Karnataka State Open University  
Mukthagangotri, Mysuru-570006

**Prof. M. Mahadevi**

**Convener**

Dean (Academic)  
Karnataka State Open University  
Mukthagangotri, Mysuru-570006

### COURSE COORDINATOR

**Shilpa Rani N R**

Chairperson

Department of Studies in Library and Information Science  
Karnataka State Open University, Mukthagangotri, Mysuru-570006

### COURSE EDITORS

**Prof. M A Gopinath**

Professor (Retd.) in LISc  
DRTC, ISI Building, Mysore Road,  
Bangalore

**Prof. A Y Asudi**

Professor (Retd.) in LISc  
Bangalore University  
Bangalore

**Dr. N. S Harinarayana**

Senior Lecturer  
Dept. of Library & Information Science  
University of Mysore, Mysore -06

**Prof. V. G. Talwar**

Professor in LISc  
Dept. of Library & Information Science  
University of Mysore, Mysore -06

### COURSE WRITER

**Prof. K S Raghavan**

Head, Library & Information Science  
DRTC, Bangalore

### BLOCK EDITOR

**Prof. N B Pangannaya**

Retd. Professor of LISc,  
University of Mysore, Mysore -06

### PUBLISHER

**Registrar**

Karnataka State Open University  
Mukthagangotri, Mysuru-570006

Developed by Academic Section KSOU, Mysore

**Copy Right: KARNATAKA STATE OPEN UNIVERSITY, 2017**

© All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from Karnataka State Open University.

This courseware is printed and published by The Registrar, Karnataka State Open University, Mysuru for limited use only. No individual or collaborative institution can use / print / distribute in any form without the written permission from KSOU. For user rights of this content and for other queries contact The Planning and Development Officer, KSOU, Mysuru 570 006.

Digital delivery of this courseware is also available for those who opt. For more details visit

[www.ksoustudymaterial.com](http://www.ksoustudymaterial.com) or [www.ksoumysore.edu.in/digitalcontent](http://www.ksoumysore.edu.in/digitalcontent)

---

**M.Lib.I.Sc – 5: Information Systems: Architecture and Retrieval**  
**Block-4: Models of Information Retrieval**

---

**Block Introduction**

Information retrieval is vital for organization with store of documents (convention and non-conventional) so as to facilitate finding of relevant information quickly and easily by the user. In searching, we are enabling users to define what they are interested in and for relevant information to be extracted from a sea of data. At this stage of searching an Information Retrieval System, users' queries are received and interpreted, appropriate search statements are formulated before making the actual search either in a traditional or modern electronic library. Searching process consists of matching queries with the document profile or database and is performed manually or through computers (batch mode and or /on-line) at local/national/global levels). In Computerized information retrieval systems, the judgment as to whether a document is relevant or not to a given query is based on the topicality or lexical similarity between the query terms and the document terms or document space (organized set of documents). Various models have been proposed to represent information retrieval systems and procedures. The units in this block will familiarize the students with these models of Information Retrieval.

Unit 13 explains the retrieval process along with a diagrammatic representation and the parameters involved in information retrieval. This leads to develop models in IR. IR models are presented in a diagram to enable the students to have a generalized view of IR models.

Unit – 13 provides a brief account of Boolean model, its application in search formulations and its advantages and disadvantages affecting the retrieval.

Probability thereby has been used as a means for modeling the retrieval process in mathematical terms. Probabilistic approaches attempt to estimate or calculate in some ways, the probability that a document will be relevant for a particular use. Unit 14

describes the basic terms of probabilistic approach to retrieval and the prevalent IR models such as Vector model, Inference Network Model, and Belief Network Model.

Unit 15 explains cognitive models and application of expert system (Artificial Intelligence) in Information Retrieval. Cognitive models allow us to predict and help prepare appropriate responses to combat or help users of Information Retrieval Systems. Cognitive models (behaviour psychology) have developed methods for systematically recording observations that can be statistically analyzed. A holistic approach is used to explain the cognitive models of IR. You are given a succinct understanding of Algebraic models, the origin of cognitive models, and the issues underlying the design of user models and their advantages and limitations. Development of expert system is explained taking CANSEARCH and PLEXUS as examples. Other experiments employing expert system for IR are also indicated in this unit.

Hypertext is the creation and representation of links between discrete pieces of digital data (text, video, audio, multimedia, etc.) Any automated retrieval system that enables users to access quantities of text called hypertext system. It includes a database structure in the form of a network and a retrieval mechanism that allows navigation and browsing. Hypertext and hypermedia are extensively used in web based information retrieval systems. The unit also presents the deference between Hypertext model and other models of IR and also certain important issues associated with hypertext model. Unit 16 discusses the basic features, components and the advantages of hypertext retrieval model. The advances in the area of Artificial Intelligence (AI) and expert systems have indicated the possibility of developing information retrieval systems that can behave intelligently (Intelligent Information Retrieval Systems). This unit discusses very briefly the area of AI and its application in retrieval process and some initiatives in this direction.

**Prof. N B Pangannaya**

## UNIT – 13

### PHILOSOPHICAL APPROACHES, EMPIRICAL MODEL, STATISTICAL MODELS PROBABILISTIC MODELS

---

13.0 Objectives

13.1 Introduction

13.2 Modeling Information Retrieval

13.3 Classification of IR Models

13.4 Classical Models of Information Retrieval

13.5 The Boolean Model

13.6 Vector Space Model

13.7 Probabilistic Model

13.8 Other Probabilistic Approaches to IR

13.9 References and Further Readings

13.10 Summary

13.11 Answers to Self Check Exercises

13.12 Key Words

13.13 References and Further Reading

#### **13.0 Objectives**

On completion of this unit you will develop an adequate understanding of:

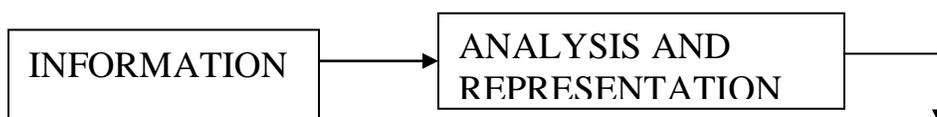
- The notion of '*modeling*' in information retrieval and what it involves
- The different approaches to Empirical Model of Information Retrieval
- The Characteristics features of the different philosophical approaches and their principal limitations
- On reading this Unit you would be in a position to understand the concept and characteristics of - The Boolean Model; Vector Space Model; Probabilistic Model and Other Probabilistic Approaches to IR

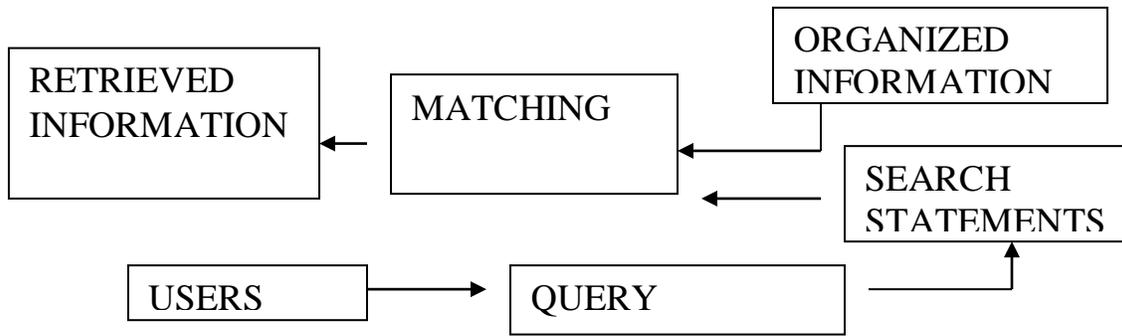
#### **13.1 Introduction**

Information Retrieval is an interactive process. By its very nature it is an uncertain process. Some of these uncertainties are likely to remain so. There are several factors that contribute to this:

- The notion of '*Relevance*' which is central to information retrieval has been and continues to remain a highly subjective notion
- There appears to be no consensus on what '*Information*' is in the context of Information retrieval
- The number of parameters involved in *Information Retrieval* is indeed large

The process of Information Retrieval (**IR**) gets initiated when a person recognizes a gap in his / her '*information store*' in the context of a '*problem situation*' / '*task to be accomplished*', etc and seeks relevant information to fill this gap. The purposes for which information is sought are many and varied ranging from mere curiosity to research to decision-making and problem solving, etc. By usage, however, the term has a much broader connotation than the mere process of information retrieval and is often used to include most of the activities that are carried out to facilitate information retrieval. In a broad sense IR is concerned with providing effective access to items of information via some mechanism for representation and organization of information items. Ideally IR should result in retrieving all those information objects (documents) that satisfy the needs of the user and no document that is not relevant to the needs of the user. These two desirable features of an IR System are the bases for developing the two parameters for evaluation of information retrieval systems, viz., Recall and Precision. Before an IR process can begin, it is necessary for the user to specify his / her information requirements. This, normally expressed in the form of a query in a natural language, is transformed into a search expression, which can be processed by the search engine of the information retrieval system. A diagrammatic representation of the Information retrieval process would appear as:





**Fig.1 The IR Process**

(Source: Chowdhury, G.G. Introduction to modern information retrieval. – London: Library Association, 1999. – p.4)

**Self-Check Exercise**

**1. What is IR Process?**

**Note:**

- i). Write your answer in the space given below.**
- ii). Check your answer with the answers given at the end of this Unit.**

.....

**13.3 Modeling Information Retrieval**

It is a good idea to begin by understanding what a ‘model’ is supposed to be. *The Compact Oxford English Dictionary* defines ‘model’ (noun) as a three-dimensional representation of a person or thing, typically on a smaller scale. ‘Model’ is also used to refer to a simplified mathematical description of a system or process, used to assist calculations and predictions. A model is an embodiment of the theory in which we define a set of objects about which assertions can be made and restrict the ways in which classes of objects can interact. A retrieval model specifies the representations used for documents and information needs, and how they are compared. (Turtle & Croft, 1992) According to Baeza-Yates & Ribeiro-Neto an information retrieval model is a quadruple [D,Q,F,R(qi,dj)] where

- D is a set of representations for the documents in the collection
- Q is a set of representations for the user information needs (queries)

- F is a framework for modeling document representations, queries, and their relationships
- $R(q_i, d_j)$  is a ranking function which associates a real number with a query and document representation (Baeza-Yates & Ribeiro-Neto, 1999)

An 'IR Model' for the purpose of our understanding is therefore, essentially a representation of the process of information retrieval in its broadest sense. Information retrieval as should be evident involves a number of objects and processes and an IR model can therefore be perceived as a theoretical framework within which we can define and make statements about these objects and processes involved in IR. There are broadly three major components in IR.

- Representations / surrogates of information resources having the potential to meet the information needs of users of the information retrieval system
- Representations of the information needs of users (search expressions)
- The process of matching these two representations

Any IR model should, thus, specify the representations used for information resources, user needs, and the process of matching these, which is primarily an exercise in matching the similarities between the two representations to retrieve information. Clearly the central problem in this relates to judging which documents in the system are relevant to a given query and which are not.

**Self-Check Exercise**

**2. What is Information Retrieval Model?**

**Note:**

- i). Write your answer in the space given below.**
- ii). Check your answer with the answers given at the end of this Unit.**

.....

**13.3 Classification of IR Models**

The Classification of Information Retrieval Models as already mentioned there are a large number of factors and parameters involved in Information retrieval. Besides, there are certain major issues that relate to document and query representations. For example,

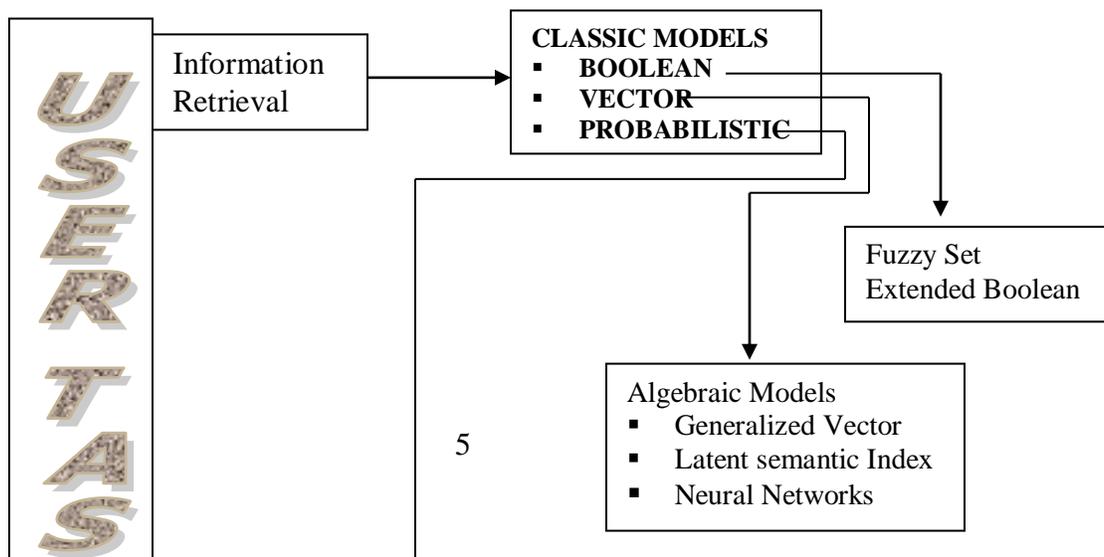
document representations and query specifications are at best only approximate semantic representations of the subject content of a document or user's actual information need. There are indeed a very large number of parameters that are involved in the IR process making rigorous application of logic to model IR difficult. However, there are several advantages in making efforts at developing models of IR:

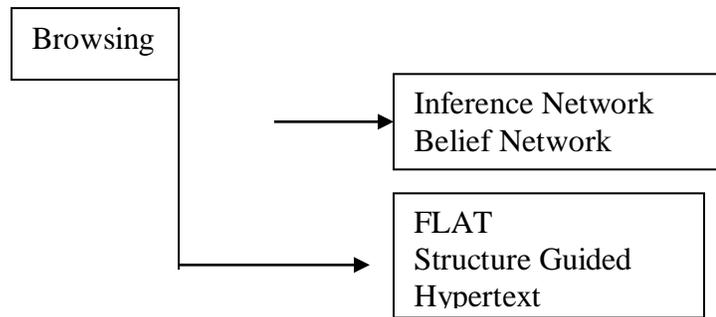
- Formalization of the different aspects of IR can help derive abstract models independent of any particular setting; this can lead to development of a theoretical framework. For IR
- Models are useful as they provide viewpoints that can help build more efficient computer-based IRs.
- They can help in the process of comparing the relative merits and limitations of different approaches to IR
- Above all efforts at modeling IR help clarify basic concepts of IR leading to a more scientific approach to its study

An examination of the relevant literature suggests that there are four main directions of research in modeling IR. These are:

- Logical Approach
- Probabilistic approach
- Vector Space Approach
- Cognitive Approach

It must be noted here that research in cognitive approach has still been of a very limited nature and covers only certain (not all) aspects of IR. A diagrammatic representation of models of IR can be found below;





**Fig. 2 Overview of Models of IR**

The diagram suggests that there are basically two types of tasks for which users approach Information Retrieval Systems; to browse records / items with the hope of finding something useful and relevant and / or to identify resources relevant to a task / problem on hand (Information Retrieval). By definition the latter task has to be more focused. In a conventional information retrieval system such as an OPAC, the documents or their surrogates (bibliographic records) remain relatively static while new queries are submitted to the system. This is what happens in a search of OPAC or any other bibliographic database. As against this there are systems that perform filtering tasks in which the users' queries remain relatively static while new documents enter the system. For example, this is precisely what happens in a SDI system in which incoming documents are matched against a relatively static user profile to identify documents relevant to the user. In the first kind of situation the system emphasizes ranking the documents on the basis of their relative degree of relevance to the user's query. In the second kind of situation the emphasis is on building a profile of the user, which is as close an approximation as possible to the user's information needs and requirements. While such systems designed to perform the task of filtering do not normally present the user with a ranked output, it is perfectly feasible to build into a filtering system some mechanism of attaching '*weights*' to documents and leave out those with a ranking below a pre-defined threshold. It should therefore be clear that filtering is also a conventional information retrieval task. It is therefore logical to examine IR models in terms of their

feature that it employs as the basis to rank documents. This should lead us to a generalized view of Information Retrieval models: An Information Retrieval model is a quadruple of logical representations for the resources in a system (collection), logical representations of users information requirements (queries), a framework for modeling the above two and a ranking mechanism that assigns a weight to a document vis-à-vis a query. Against this background let us discuss briefly the different models of information retrieval.

**Self-Check Exercise**

**3. What are the advantages of Information Retrieval Models?**

**Note:**

**i). Write your answer in the space given below.**

**ii). Check your answer with the answers given at the end of this Unit.**

.....

.....

**13. 4. The Classic Models of Information Retrieval (IR)**

There are three classic models of IR, Boolean Model, Vector Model and Probabilistic Model. All classic IR models view that a document representation which is in effect a description of a document consists of a set of descriptors / keywords / classification codes or a combination of these. These index terms / access points are perceived to be complete or partial semantic representations of the main themes (*'aboutness'*) of the document. At one extreme the logical view of a document could be the full text of the document (e.g. many web search engines consider all unique words in a document as index terms). On the other hand the logical view could be a limited number of index terms extracted from or assigned to a document based on some understanding / logic of how representative these are of the theme of the document. In either case it is important to recognize that in reality, given a set of index terms to describe a document, all of them are not equally

significant; some of them are necessarily less important than some others in terms of their representative character. In other words each index term carries a certain weight that quantifies the importance of the index term in describing the ‘*aboutness*’ (the semantic content) of the document.

**13.5 STATISTICAL MODELS** The goal of Information Retrieval (IR) is to provide users with those documents that will satisfy their information need. We use the word “document” as a general term that could also include non-textual information, such as multimedia objects. Users have to formulate their information need in a form that can be understood by the retrieval mechanism. The contents of large document collections need to be described in a form that allows the retrieval mechanism to identify the potentially relevant documents quickly. In both cases, information may be lost in the transformation process leading to a computer usable representation. Hence, the matching process is inherently imperfect.

Information seeking is a form of problem solving. It proceeds to the interaction among eight sub processes:

1. Problem recognition and acceptance,
2. Problem of definition
3. Search system selection
4. Query formulation
5. Query execution
6. Examination of results (including relevance feedback)
7. Information extraction and
8. Reflection / Termination

### **13.6 The Boolean Model**

The Boolean model is a fairly simple retrieval model and is based on the concept of ‘*sets*’. A descriptor / keyword\* defines a set of documents; i.e. all the documents that are assigned the same descriptor belong to the *set* defined by that descriptor. As any

---

\* Technically the terms ‘*descriptor*’ and ‘*keyword*’ are not synonymous; however in this learning package unless otherwise mentioned they will be used as though they are synonyms.

document can be assigned multiple descriptors (depending on the number of concepts it discusses), usually most documents in an information system are members of two or more sets; i.e. a document is a member of all those sets defined by the descriptors assigned to it / keywords present in it. Users' queries are also specified as Boolean Expressions that have reasonably precise semantics. Because of its simplicity it was widely adopted by many of the commercial bibliographic information retrieval systems. Some of the important features of the Boolean model which also define its limitations are:

- It is an exact match system;
- Document attributes are assigned binary values {0,1}
  - Document attributes can include not only keywords / descriptors but also more complex attributes such as dates, source, authors, language, etc.
- A user's query is specified as a Boolean expression using attribute values related by the operators - AND, OR, NOT
- All documents that match the search specification are treated as equally relevant and retrieved

The Boolean model is widely used in commercial information retrieval systems beginning with the online systems of 1970s. The Boolean search is generally based on an inverted index file. A brief explanation of the Boolean operators and their application is in order. As already mentioned there are basically three Boolean relational operators: **AND, OR, and NOT**

A search using the Boolean OR operator (e.g. *Library OR Libraries; Information OR Data*) considers a document relevant if the document contains at least one of the terms specified in the query. In other words, use of Boolean OR operator between two search terms in a search expression will result in the retrieval of all those documents which contain either of the terms linked with an OR as also documents that contain both the terms so linked. Ideally the OR operator is used to connect synonyms, near synonyms, etc. In practice the OR operator is also often used to link a broader term with its narrower terms. For example, if a user is interested in retrieving documents related to 'India', it is

likely that the user will also be interested on material that may deal with specific regions, states, etc of *India*. The OR operator could be used in the search expression to link *India* with the names of all the regions and states.

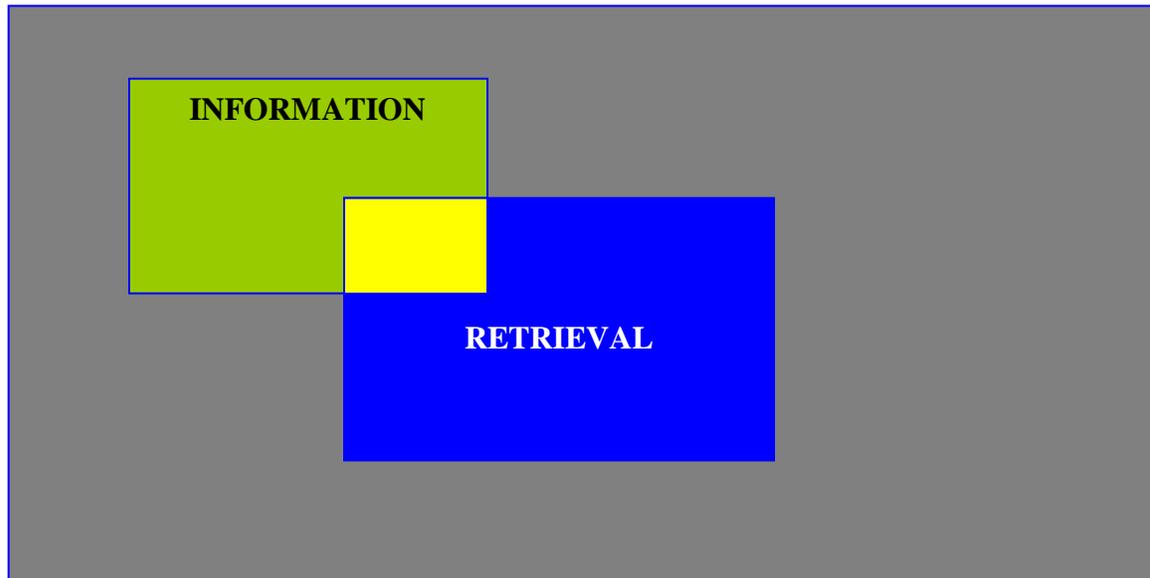
A search using the Boolean AND operator (e.g. *Information AND Retrieval*) considers a document relevant only if all the terms connected using the AND operator are present in the document. The Boolean AND operator is used in a search expression to connect the different facets of a subject on which information is required. For example, if a user is interested in ‘History of India’, the search expression, *India AND History* is used to retrieve the relevant documents.

A search using the Boolean NOT operator (e.g. *South Asia NOT India*) will consider as relevant only those documents in which the term ‘*South Asia*’ is present and ‘*India*’ is not present. In other words the search would consider as irrelevant any document in which the term ‘*India*’ is present even if the document were to contain the search term ‘*South Asia*’. The Boolean operators **OR**, **AND**, and **NOT** correspond to the set operations: *set union*, *set intersection* and *set disjunction*. A diagrammatic representation of the Boolean operators is given below:

Examples: INFORMATION **OR** DATA

INFORMATION **AND** RETRIEVAL

INFORMATION **NOT** DATA



When the two search terms are linked using the Boolean AND only those documents (Represented by the area painted yellow) will be retrieved. The use of OR to link the two terms will result in the retrieval of documents corresponding to both the rectangles including the yellow portion; the use of the search expression ‘Information NOT Retrieval’ will result in retrieving only those documents represented by the area in green. The OR operator is used when one is searching for documents which have either or both the terms linked this way. The use of OR would be appropriate to link two or more synonyms or near synonyms. The OR operator expands a search and treats the operands as equivalent. AND and NOT on the other hand, limit a search. One and the same search expression may also involve multiple uses of one or more of the Boolean operators.

### **13.6.1 Limitations of the Boolean Model**

Let us examine in some detail some of the major limitations of the Boolean Model. The Boolean models are limited in their scope. The models are based on a set of premises, not all of which is valid:

- In a Boolean search expression all the terms are assumed to be of equal weight. It does not recognize the need for or make a provision for assigning a degree of importance to terms in a query. Conventional Boolean models adopt a kind of binary classification in which documents are considered either relevant or irrelevant. Such an

approach has limited value since for users the notion of relevancy is not static but dynamic; i.e., for a given user, what is irrelevant today may be relevant tomorrow and vice versa. There is also the valid concept of degree of relevance of a document to a query, which is completely overlooked

- For a search expression involving the OR operator between search terms, documents containing any one of the query terms are treated as relevant as documents containing all of the query terms
- For a search expression involving the AND operator a document not containing just one of the terms is treated as irrelevant as documents not containing any of the terms
- A search expression involving the NOT operator eliminates a document even if the document casually uses the term connected by the OR operator
  - The output in a Boolean search is not ranked and this means that the users may have to scan the entire set of retrieved documents

To overcome some of these limitations of the Boolean model of information retrieval, Gerald Salton, Edward Fox and H. Wu introduced the Extended Boolean IR Model in 1983\*. This is an extension of the Boolean model in which partial matches (which are either completely rejected as irrelevant or assumed to be as relevant as any other in the classic Boolean IR model) are interpreted as Euclidean distances in a vector space. The major limitations of the Boolean model derive from the fact that the model does not provide for term weights and for ranking the output. The Extended Boolean model, while it considers all the terms in a query, is flexible and provides for adjusting the rigor with which the operators, AND and OR have to be applied in a given search. This is done via a p-value which adjusts the strictness of the operators AND and OR. The model is based on computing the similarity between a document and the query by measuring the distance from **point 1** for AND (which is indicative of a position in space in which all of the query terms are present), and from **point 0** (none of the query terms present) for operator OR. It is expected that the p-value for a particular search is specified at the time of query. In a way since the Extended Boolean model introduces the notion of distances, it is no longer strictly a model based entirely on set theory as the conventional Boolean Model is.

---

\* G. Salton, E. Fox, and H. Wu. Extended Boolean Information Retrieval. *Communications of the ACM*, 1983, 26(11): 1022-1036

**Self-Check Exercises**

- 1. What is Boolean Model of IR?**
- 2. What are the limitations of Boolean IR Model?**

**Note:**

- i). Write your answer in the space given below.**
- ii). Check your answer with the answers given at the end of this Unit.**

.....  
.....

**13.7 The Vector Model:** The Vector model is a classic model of information retrieval that is based on representing documents and queries as vectors of terms. It was mentioned in the preceding section that the use of binary weights (0 or 1) limits the utility of the Boolean model in that the user ends up retrieving either too few or too many documents. In other words the classic Boolean model lacks the required discriminatory ability and grading scale expected of a good information retrieval model. Use of such binary weights is probably perfectly suitable for data retrieval, but not for information retrieval in which there is always a certain amount of uncertainty. It is therefore difficult to translate an information need into a Boolean expression that requires precise semantics. The Vector model seeks to overcome some of these limitations by assigning non-binary weights to query terms. These weights are employed by the retrieval system to compute the degree of similarity between a document and the user’s query and sort the retrieved documents in the order of decreasing degree of similarity. Thus the user is presented with an output that includes documents that could only be partial matches to his/her information need and allow the user to scan the output and decide on a threshold for cut off. Such an approach is more in tune with the requirements of information retrieval. Conceptually, in the Vector model words in a query form a subspace over which the document space is projected. All documents projecting in or near the subspace are deemed to be potentially relevant to the query.

In this model documents and users queries are represented as vectors (the number of dimensions in the vector being the number of descriptors or keywords in the document or the number of terms in the search expression). Suppose we assume that the vector **D**

represents a document and the vector  $\mathbf{Q}$  represents a query (search), the Vector IR model computes the degree of similarity between the document and the query as the correlation between these two vectors. This correlation can also be quantified as the Cosine of the angle between the two vectors. Unlike the classic Boolean model, which categorizes every document as either '*relevant*' or '*not relevant*' and outputs only those that are considered relevant, the Vector space model merely ranks the documents according to their degree of similarity to the query. One can establish a threshold of the minimum degree of similarity acceptable to consider a document relevant and retrieve only those documents above this threshold.

It should be obvious that an important requirement for the Vector model of IR is a mechanism for computing weights of document terms. There are a number of different document-term weighting techniques. A detailed discussion of these techniques is not required here. However, it is important to understand the basic principle that is employed for assigning weights to document terms. The technique employed for this purpose views information retrieval as a clustering problem. Given a query, the problem of clustering documents in a collection is to divide the documents into two broad classes of objects: the set of objects that have the required degree of similarity to the query and those that do not meet the required degree of similarity. Suppose we were to consider that the subset  $\mathbf{S}$  in a collection of documents meets the required degree of similarity we must specify the attributes of these documents to be included in the set  $\mathbf{S}$ , which distinguish these documents from those that are to remain outside the set  $\mathbf{S}$ . The first one requires quantifying **intra-cluster similarity** and the second one requires quantifying **inter-cluster dissimilarity**. For computing both these, Vector models generally employ term frequency counts.

**The principal advantages of the Vector model are:**

- The retrieved documents are sorted and ranked on the basis of the degree of similarity between the document and the query thus positioning the supposedly most relevant documents at the top of the list of retrieved documents

- Its approach of using term weights results in better retrieval performance

### 13.7.1 Generalized Vector Space Model

The classic Boolean and Vector models assume that index terms are independent of each other and are not correlated. In reality however, this is not correct as index terms are in fact correlated. The generalized Vector Space model takes this factor into account. This model (which is an extension of the Vector model) uses the idea that co-occurrence of index terms in a document is an indicator of correlation among them. However, whether usage of term dependencies will improve retrieval performance is still disputed. In addition the model is more complex and involves some expensive computation compared to the classic model. A detailed discussion of the model is not feasible here.

#### Self-Check Exercise

#### 3. What are the advantages of Vector Model?

**Note:**

**i). Write your answer in the space given below.**

**ii). Check your answer with the answers given at the end of this Unit.**

.....

.....

### 13.8 Probabilistic Models

Maron and Kuhns had proposed a probabilistic retrieval model as early as in 1960\*. The model is based on probability theory. In essence it is argued that the probability that a particular document would be relevant to a user can be assessed by a calculation of the probability, for each document in the collection. S.E. Robertson and K. Sparck Jones introduced a probabilistic model in 1976.\* As the name suggests the model seeks to view the problem of information retrieval

---

\* M. E. Maron and J.L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7, 1960, p.216-244

• S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *JASIS*, 27(3) 1976; p. 129-146

within a probabilistic framework. The basic idea begins by suggesting that for any query in an information retrieval situation, the document collection can be imagined to have a document set containing only and exactly the relevant documents and no irrelevant document. This set of documents represents the *perfect set*. It is not difficult to visualize that an Information Retrieval system will contain documents that answer a particular query. The fact that all these documents answer a particular query in turn requires that they all share certain common attributes. The problem of information retrieval can now be seen as a process of specifying in a query those attributes that all the documents of this *perfect set* of relevant documents share.

Information Retrieval Systems employ keywords / descriptors (either those derived from the text being described or those assigned by an indexer / searcher) to represent the semantics of the document / query. Essentially the semantics of the descriptors / keywords assigned to a document are expected to specify the attributes of the document that assist determination of whether a given document should belong to the *perfect set* (of relevant documents) or remain outside that set in the context of a given query. It is important to remember that these properties of the documents are unknown at the time of querying. In other words at the time of querying an IR system we are generally guessing what these attributes could be based on the semantics of the query. This guess is the initial probabilistic description of the attributes of the *perfect set*. An interaction with the end user is normally initiated with the idea of enhancing and improving this probabilistic description. The end user usually examines the retrieved documents to decide on their relevance. This feedback is used as the input for refining the probabilistic description; i.e. an effort to make it a closer approximation of the attributes of the *perfect set*. In practice this exercise is repeated many times before the final probabilistic description is arrived at.

**The principal premises on which the Probabilistic Model of IR is grounded are:**

- It is possible to estimate the probability that a given document  $D$  in an IR System is relevant to the query  $Q$ ;
- This probability is a function of the concerned document and query representations;

- There is, for every query, a subset of the documents in the IR system which is best acceptable to the user as relevant; the documents that are not a part of this subset are not relevant to the Query
- Retrieving this subset of documents maximizes the overall probability of relevance to the user
- An Index term / Query term can only take one of the two values – *Zero* or *One*

At the time a query is specified and a search is yet to be made, there are no retrieved documents. The basic instrument in any information retrieval for separating relevant documents from those that are not relevant is the matching function. The probabilistic model uses probability theory to define this matching function. In the probabilistic model of IR the matching function is derived on the basis of some knowledge of the distribution of index terms among the documents that constitute the collection of the IR system (or a subset of documents in the collection chosen using an appropriate technique such as sampling, trial retrieval, etc). The data / information gathered from these documents (retrieved, say, in a trial retrieval) is used to set the values of certain parameters associated with the matching function.

### **13.8.1 Estimating Relevance**

The relevance of a document can be fully ascertained only after the user has gone through the text of the document. However, in IR we are faced with the issue of guessing the relevance of a document for a particular query. Relevance is guessed on the basis of document description (metadata about the document that is available) and in some cases, on the basis of how it is related to other documents. A way of quantifying this guess is to estimate the probability of relevance of a document in the collection to a given query ( $P_Q$ ). One type of data about a document that IR systems generally possess that can assist estimating its probability of relevance to a given query is the term occurrence frequency data. Given the probability of relevance of a document to a given query it is possible to rank the documents in a collection in a sequence of decreasing probability of relevance to the query. However, this ranking assumes that the probability of relevance can not only

be calculated but also calculated accurately. In reality, in the IR situation given a query specification, one is not aware of:

- Which of the documents in a collection are relevant?
- How many relevant documents are there in the collection?

In real life situation the best that can be done is to guess  $P_Q$  based on a trial retrieval and to improve this guess by iteration. Different approaches have been suggested for determining the probability of relevance. Let us briefly examine the three different approaches to probabilistic retrieval:

Maron and Kuhns suggest that, the probability that a document  $D$  will be judged relevant by a user submitting a query  $Q$  consisting of term  $T$  is the ratio of the users who submit the same query term and consider the document  $D$  relevant to the total number of users who submit the query term  $T$ . This approach requires that historical information be used for determining probability of relevance, i.e. prior knowledge of how many of the users who had submitted the query term  $T$  had judged the document  $D$  relevant.

The approach suggested by Robertson and Sparck Jones is different from the above. The substance of this approach is that probability of relevance can be calculated for a set of documents sharing a particular attribute in relation to a given user rather than for a set of users giving the same query term. It is suggested that given a set of documents possessing a particular attribute and a given user the probability that the user will judge a document relevant is the ratio of the number of documents that have the attribute to the total number of documents.

Salton and McGill have suggested a third approach. The substance of this approach is that if estimates for probability of occurrence of various terms in relevant documents can be calculated, then the probability that a relevant document (or a non-relevant) will be retrieved can be estimated.

While many experiments have indicated that the probabilistic approach yields satisfactory results, the results have not been sufficiently better than those obtained using the conventional Boolean approach. The probabilistic approach is premised on two important parameters or measures:

- The probability of Relevance -  $P_r$

- The probability of Non-relevance –  $P_{nr}$

When one considers relevance as a binary property (i.e. either a document is relevant or not relevant),  $P_{nr} = 1 - P_r$

There are two major cost parameters associated with the approach; the cost of retrieving an irrelevant document and the cost of missing (non-retrieval) a relevant document. The principal advantage of the probabilistic approach is that the retrieved documents can be ranked

**Self-Check Exercise**

**4 Briefly explain the Probabilistic Model.**

**Note:**

- i). Write your answer in the space given below.
- ii). Check your answer with the answers given at the end of this Unit.

.....  
 .....

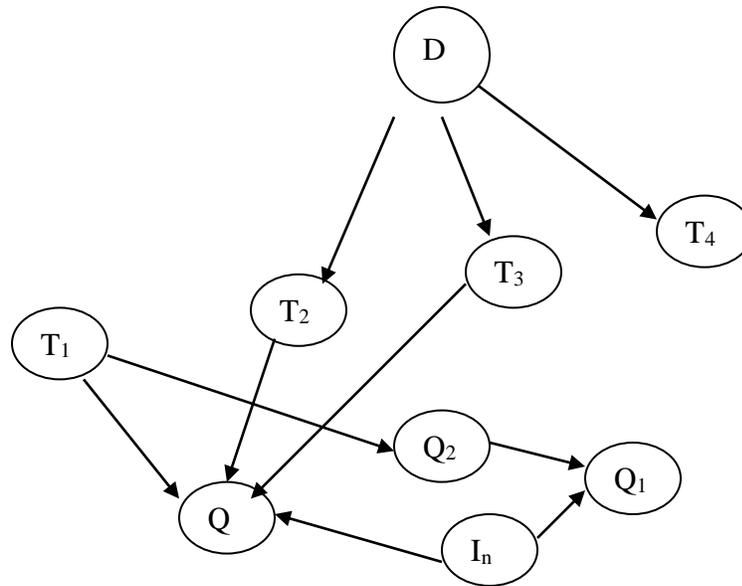
**13.9 Other Probabilistic Approaches to IR**

Attempts have also been made by researchers to develop some alternative probabilistic approaches to IR. Two such models are based on Bayesian networks. Bayesian networks are essentially directed acyclic graphs, the nodes of the graphs representing random variables. The links between these nodes indicate causal relationships between the nodes and the strength of these relationships are expressed as conditional probabilities. Bayesian networks are considered useful in that they are able to provide a formal basis for ranking a document, which can be used to improve retrieval performance. **Basically two models have emerged out of Bayesian networks; viz., the Inference Network Model and the Belief Network Model.**

**13.9.1 Inference Network Model**

The Inference Network Model adopts an epistemological approach to IR. Random variables are associated with index terms, documents and users' queries. It is assumed

that documents are being observed in the search for relevant documents. Documents, Index terms and query are all represented as nodes in the network.



**Figure 4: Basic Inference network Model**

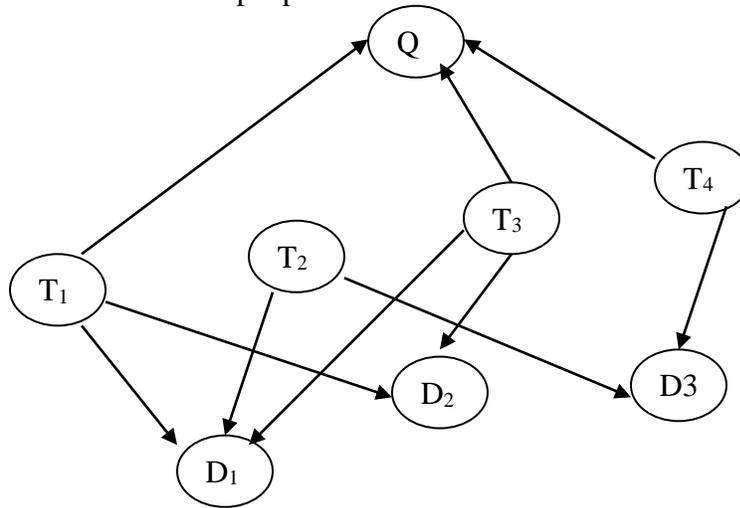
(Source: Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

*Modern Information Retrieval*. -- Delhi: Pearson Education, 2004. -- p. 50)

The above figure illustrates an inference network for IR. The document  $D$  has been assigned three index terms, viz.,  $T_2$ ,  $T_3$  and  $T_4$ . The user's information need  $I_n$  expressed as query  $Q$  has the query terms  $T_1$ ,  $T_2$  and  $T_3$ . The nodes  $Q_1$  and  $Q_2$  represent alternative query formulation for the query  $Q$  based on additional information that may be available. The rank of the document  $D$  with regard to the query  $Q$  is a measure of the amount of evidential support the observation of  $D$  provides to  $Q$ .

### 13.9.2 Belief Network Model

The **Belief Network Model** differs from the Inference Network Model in that it adopts a clearly defined sample space. Because of this, the network topology is different and separates the document and query portions of the network.  $K$  is the set of all index terms and is viewed as concept space.



**Figure 5: Basic Belief Network Model**

(Source: Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. -- Delhi: Pearson Education, 2004. -- p. 58)

Just as in the Inference Network Model a query  $Q$  is viewed as composed of index terms. The difference is that documents are also treated similarly as made up of index terms. This is the topological difference between the two models. The ranking of a document in relation to a given query is seen as a concept matching relationship.

While there is a great deal of similarity between the two approaches, the Belief Network Model is more general from a theoretic point of view. Bayesian network models are in effect variants of the probabilistic model that allow the use of evidences to support the fixing of relevance rank of a document. The Bayesian models, particularly the Belief Network Model can be made to incorporate feedback information from previous search sessions to enhance retrieval.

**Self-Check Exercise**

**5. What is Inference Network model and Basic Belief Network Model?**

**Note:**

- i). Write your answer in the space given below.**
- ii). Check your answer with the answers given at the end of this Unit.**

.....  
.....

**13.10 Summary**

The process of Information Retrieval (**IR**) gets initiated when a person recognizes a gap in his / her ‘*information*. There are broadly three major components in IR i.e. Representations / surrogates of information resources having the potential to meet the information needs of users of the information retrieval system, Representations of the information needs of users (search expressions) and The process of matching these two representations. There are several advantages in making efforts at developing models of IR are Formalization of the different aspects of IR can help derive abstract models independent of any particular setting; this can lead to development of a theoretical framework. For IR, Models are useful as they provide viewpoints that can help build more efficient computer-based IRs and they can help in the process of comparing the relative merits and limitations of different approaches to IR.

The Boolean model is widely used in commercial information retrieval systems beginning with the online systems of 1970s. The Boolean search is generally based on an inverted index file. A brief explanation of the Boolean operators and their application is in order. As already mentioned there are basically three Boolean relational operators: AND, OR, and NOT. The Vector model is a classic model of information retrieval that is based on representing documents and queries as vectors of terms. It was mentioned in the preceding section that the use of binary weights (0 or 1) limits the utility of the Boolean model in that the user ends up retrieving either too few or too many documents. The Probabilistic model is based on probability theory. In essence it is argued that the probability that a particular document would be relevant to

a user can be assessed by a calculation of the probability, for each document in the collection.

### **13.11 Answers to Self Check Exercises**

#### **1. What is IR Process?**

The process of Information Retrieval (**IR**) gets initiated when a person recognizes a gap in his / her '*information store*' in the context of a '*problem situation*' / '*task to be accomplished*', etc and seeks relevant information to fill this gap. In a broad sense IR is concerned with providing effective access to items of information via some mechanism for representation and organization of information items. Ideally IR should result in retrieving all those information objects (documents) that satisfy the needs of the user and no document that is not relevant to the needs of the user.

#### **2. What is Information Retrieval Model?**

An 'IR Model' for the purpose of our understanding is therefore, essentially a representation of the process of information retrieval in its broadest sense. An IR model can be perceived as a theoretical framework within which we can define and make statements about these objects and processes involved in IR. There are broadly three major components in IR.

- Representations / surrogates of information resources having the potential to meet the information needs of users of the information retrieval system
- Representations of the information needs of users (search expressions)
- The process of matching these two representations

Any IR model should, thus, specify the representations used for information resources, user needs, and the process of matching these, which is primarily an exercise in matching the similarities between the two representations to retrieve information.

#### **3. What are the advantages of Information Retrieval Models?**

The advantages of IR models are:

- Formalization of the different aspects of IR can help derive abstract models independent of any particular setting; this can lead to development of a theoretical framework. For IR
- Models are useful as they provide viewpoints that can help build more efficient computer-based IRs.
- They can help in the process of comparing the relative merits and limitations of different approaches to IR
- Above all efforts at modeling IR help clarify basic concepts of IR leading to a more scientific approach to its study

#### **4. What is Boolean Model of IR?**

The Boolean model is a fairly simple retrieval model and is based on the concept of 'sets'. It is an exact match system; Document attributes are assigned binary values {0,1}. Document attributes can include not only keywords / descriptors but also more complex attributes such as dates, source, authors, language, etc. A user's query is specified as a Boolean expression using attribute values related by the operators - AND, OR, NOT . The Boolean operators *OR*, *AND*, and *NOT* correspond to the set operations: *set union*, *set intersection* and *set disjunction*.

#### **5. What are the limitations of Boolean IR Model?**

The Boolean models are limited in their scope. The models are based on a set of premises, not all of which is valid: In a Boolean search expression all the terms are assumed to be of equal weight. It does not recognize the need for or make a provision for assigning a degree of importance to terms in a query. Conventional Boolean models adopt a kind of binary classification in which documents are considered either relevant or irrelevant. Such an approach has limited value since for users the notion of relevancy is not static but dynamic; i.e., for a given user, what is irrelevant today may be relevant tomorrow and vice versa. There is also the valid concept of degree of relevance of a document to a query, which is completely overlooked. For a search expression involving the OR operator between search terms, documents containing any one of the query terms are treated as relevant as documents containing all of the query terms. For a search expression involving the AND operator a document not containing just one of the terms is treated as irrelevant as documents not containing any of the terms. A search expression

involving the NOT operator eliminates a document even if the document casually uses the term connected by the OR operator. The output in a Boolean search is not ranked and this means that the users may have to scan the entire set of retrieved documents

#### **6. What are the advantages of Vector Model?**

The principal advantages of the Vector model are the retrieved documents are sorted and ranked on the basis of the degree of similarity between the document and the query thus positioning the supposedly most relevant documents at the top of the list of retrieved documents and Its approach of using term weights results in better retrieval performance

#### **7. Briefly explain the Probabilistic Model.**

Maron and Kuhns had proposed a probabilistic retrieval model as early as in 1960. The model is based on probability theory. In essence it is argued that the probability that a particular document would be relevant to a user can be assessed by a calculation of the probability, for each document in the collection. As the name suggests the model seeks to view the problem of information retrieval within a probabilistic framework. The principal premises on which the Probabilistic Model of IR is grounded are:

- It is possible to estimate the probability that a given document *D* in an IR System is relevant to the query *Q*;
- This probability is a function of the concerned document and query representations;
- There is, for every query, a subset of the documents in the IR system which is best acceptable to the user as relevant; the documents that are not a part of this subset are not relevant to the Query
- Retrieving this subset of documents maximizes the overall probability of relevance to the user
- An Index term / Query term can only take one of the two values – *Zero* or *One*

#### **8. What is Inference Network model and Basic Belief Network Model?**

The **Inference Network Model** adopts an epistemological approach to IR. Random variables are associated with index terms, documents and users' queries. It is assumed

that documents are being observed in the search for relevant documents. Documents, Index terms and query are all represented as nodes in the network.

The **Belief Network Model** differs from the Inference Network Model in that it adopts a clearly defined sample space. Because of this, the network topology is different and separates the document and query portions of the network

### 13.7 Key Word

**Information Retrieval:** Any system, usually involving computers, that performs information retrieval.

**Query:** The formal expression of an information need

**Proximity Operator:** A function that identifies pairs of terms satisfying specified Proximity conditions

**Vector Model:** A retrieval model based on viewing documents and queries as term or term weight vectors.

### 13.12 References and Further Readings

1. Chowdhury, G. G. Introduction to modern information retrieval. – London: Library Association Publishing, 1999 (especially chapters 8, 16 and 17)
2. Ellis, David. Progress and problems in information retrieval. London: Library Association Publishing, 1996
3. Baeza-Yates, Ricardo and Ribeiro-Neto, Berthier. Modern information retrieval. – Delhi: Pearson Education, 2004
4. Ellis, David. Progress and problems in information retrieval. London: Library Association Publishing, 1996
5. Baeza-Yates, Ricardo and Ribeiro-Neto, Berthier. Modern information retrieval. Delhi: Pearson Education, 2004

---

**UNIT – 14****COGNITIVE MODELS AND APPLICATION OF EXPERT SYSTEM IN INFORMATION RETRIEVAL: I<sup>3</sup>R, CANSEARCH, PLEXUS ETC.,**

---

- 14.0 Objectives
- 14.1 Introduction
- 14.2 Algebraic Models:
- 14.3 Cognitive Information Retrieval
- 14.4 CANSEARCH
- 14.5 PLEXUS
- 14.6 I<sup>3</sup>R (*Intelligent Intermediary for Information Retrieval*)
- 14.7 Hypertext association approach to IR
- 14.8 Hypertext Model
- 14.9 Expert Systems
- 14.10 Summary
- 14.11 Answers to Self Check Exercises
- 14.12 Key Words
- 14.13 References and Further Readings

**14.0 Objective**

On reading this Unit you would be in a position to understand the concept of – Hypertext Model and Expert System Model.

**14.0 Objective**

On reading this Unit, you would be in a position to understand the characteristics of Algebraic Models; Cognitive Information Retrieval; and application of Expert Systems in IR such as CANSEARCH; PLEXUS; I<sup>3</sup>R (*Intelligent Intermediary for Information Retrieval*)

**14.1 Introduction**

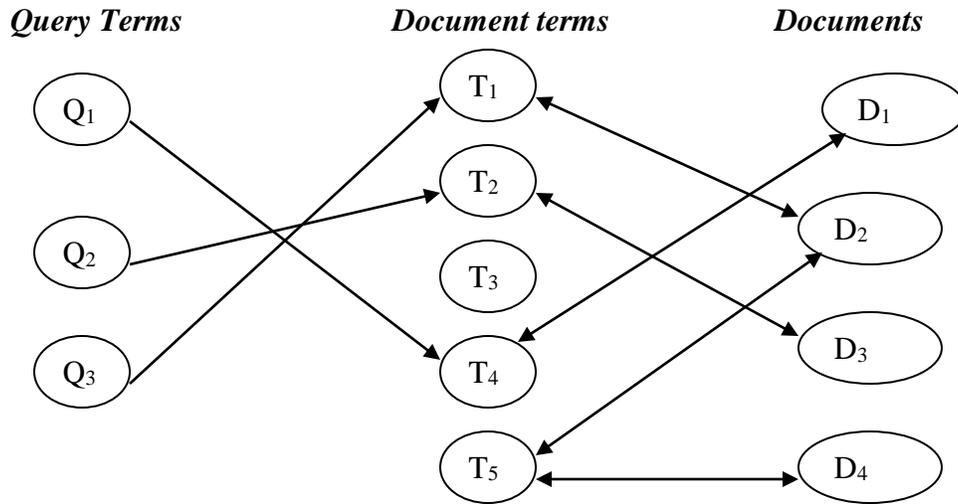
This module consists of learning materials covering four units related to IR models. What is attempted in the following pages is to provide an overview of the Cognitive IR

models. All the IR models have certain common premises. However an understanding of these is necessary to have a ‘*holistic*’ view of the subject.

## 14.2 Algebraic Models

Two other algebraic IR models are the **latent semantic indexing** and the **neural network** models. The Latent Semantic Indexing proceeds on the premise that index terms derived from texts of documents often lead to poor retrieval performance in view of the limitations of the related index term derivation processes in summarizing the contents of documents adequately and accurately. A number of limitations of natural language text in serving effectively as the basis for deriving index terms have been discussed in related literature. If the process of retrieval could be made to match a given document to a given query on the basis of the concepts rather than terms in the text, it should lead to better retrieval performance. Latent semantic indexing is an approach that seeks to address this issue. Latent Semantic Indexing is a technique based on multidimensional scaling for identifying major concepts in a document or document collection. Given a document collection (say, with  $N$  number of documents in it), it is possible to construct a term-document association matrix of  $N$  columns with  $t$  rows, where  $t$  is the number of index terms derived from the  $N$  documents in the collection. It is possible to assign a weight to every term-document pair using some weighting technique. The IR model based on Latent Semantic Indexing the effort is to map documents and a query vector into a lower dimensional space associated with concepts allowing retrieval based on concept matching rather than term matching. For instance a document could be retrieved because it shares concepts with another document that has been judged relevant to a given query.

The **Neural Network Model** seeks to adopt the approach of pattern matching used by human brain in the IR process. In other words a neural network IR model is a simplified version of the mesh of interconnected neurons in the human brain. Since the variables involved in the IR process are documents, index terms representing documents and query terms, a graphical representation of the neural network model of IR will appear like what is given in the figure below:



**Figure 6: Neural Network Model**

The figure depicts the process of IR in a neural network model. The query term nodes send signals to the document term nodes. The document term nodes in turn generate signals to document nodes. However, the process does not end here. The document nodes in turn may generate new signals and send them back to the document term nodes (notice the bi-directional arrows between document term nodes and document nodes) and the process can continue. The signals get weaker and weaker and finally the process stops. It is possible even in the neural network model to rank the output. The neural network model has not been tested extensively and it is therefore difficult to say whether the model yields superior results or not.

**Fuzzy Set Approach:** As the very name suggests, fuzzy sets are sets whose boundaries are not clearly defined. Since the boundaries of a set are not clearly defined, the membership of a set cannot be clearly defined by a ‘yes’ or ‘no’; instead there will be degree of membership. When this membership function is associated, it can take any value ranging from **0** (zero) to **1** (one), 0 corresponding to no membership in the class and 1 corresponding to full membership in the class. Any value between 0 and 1 will indicate varying degrees of membership. When applied to information retrieval, the Fuzzy set approach involves the use of a thesaurus. A thesaurus as we know it defines for a given term those terms that are hierarchically and associatively related to it. These

related terms provided in a thesaurus could also assist in the search process by facilitating expansion of a search using terms related to a query term to yield higher recall or to facilitate a more precise search to yield better precision. A thesaurus can also be used to model IR in terms of fuzzy sets. A term-term correlation matrix can be created based on their correlation (co-occurrence) in a given document collection. The correlation factor between two terms ( $T_1$  and  $T_2$ ) that co-occur in the collection can be computed using data on the number of documents that contain the terms  $T_1$  and  $T_2$ , the number of documents that contain only term  $T_1$  and the number of documents that contain only  $T_2$ . This correlation matrix can be used to define a fuzzy set associated with an index / search term in which the degree of membership of every document can be computed. A document belongs to a fuzzy set associated with a term  $T_1$  if at least one of the terms by which it is indexed is strongly related to the term  $T_1$ . The search proceeds on lines similar to the Boolean model, i.e. the user submits a Boolean-like query and the procedure to determine the relevant documents is similar to what happens in the Boolean model except that one is dealing with a Fuzzy set. This allows the retrieved documents to be ranked in relation to the query. Going by the literature on the subject, Fuzzy set models do not appear to be very popular among the members of the information retrieval community as most of the discussion on the subject appears in literature on Fuzzy theory.

**Self-Check Exercise**

**1. What are Algebraic Models of IR?**

**Note:**

- i). Write your answer in the space given below.**
- ii). Check your answer with the answers given at the end of this Unit.**

.....  
.....

**14.3 Cognitive Information Retrieval**

Most IR models that we have examined so far are premised on the assumption that users approach information retrieval systems with well-defined information needs expressed as queries. In reality however, this is not the case. Most users who approach information retrieval systems generally find it difficult to express their information needs in the form

of a well-formulated query. The users' requests are, more often than not, generally not in a form that will prove effective in a search of a retrieval system. It was therefore thought that it would be beneficial to model a retrieval system based on user information needs rather than on pure query matching. This approach is based on the premise that '*information need*' of a user is dynamic and not static. This is to suggest that during the process of information retrieval, which is heuristic in nature the user should be allowed to modify and refine his '*information need*'. This constitutes the basis for the cognitive model of IR. Cognitive IR models are a comparatively recent development. This probably answers the question: why is there no established research tradition in cognitive IR compared with the more established approaches such as the Probabilistic and Boolean models? The approach seeks to use the cognitive structure of the user to build a model of the user in the system.

The origins of the cognitive model of IR can be seen in the '*Anomalous state of knowledge*' (ASK) hypothesis of Belkin. It is therefore useful to have a broad understanding of the rationale behind ASK. The ASK notion can be discussed and explained only in a cognitive framework. Let us imagine a scholarly communication situation. When an author writes a paper or some other document containing scholarly information, the text of the document is essentially a transformed representation of the cognitive knowledge structure of the individual author. When another individual studies / reads this document his cognitive structure interacts with the cognitive structure of the author of the document resulting in a transformation of the cognitive knowledge structure of recipient of the communication. The cognitive knowledge structure of an individual in respect of an entity at any particular point of time is his / her worldview of that entity. This is not static and keeps on changing as the individual receives communication related to the entity. Thus we see that there are two related aspects to the ASK framework:

- The factors that relate to a decision by an individual to communicate an aspect of his / her knowledge; the factors will include, among other things the author's perception of the state of knowledge the potential users of his communication. All these will influence the text of the communication

- The factors that relate to a decision by a user to seek information on a subject and his / her subsequent decision that a particular text would be useful in meeting this information need.

It is generally assumed that a searcher seeks information when he / she recognizes an anomaly in his / her knowledge structure related to a particular theme / entity. The purpose of seeking information is to resolve this anomaly. The process of information retrieval and the subsequent reading of texts (which are usually repeated until the user is satisfied) is therefore largely a process of resolving the anomaly. Typically the process of information retrieval involves the following steps when viewed from a cognitive point of view:

- A user initiates a search in a system recognizing an anomaly in his cognitive structure in respect of an entity (subject)
- This is converted into and expressed in the form of a request (query) and submitted to a retrieval system
- The retrieved texts are examined by the user which in essence means that the cognitive structure of the communicator (author) of a retrieved text interacts with the cognitive structure of the user
- If the recipient feels that the anomaly in his / her knowledge structure has been resolved, the process is brought to a stop; otherwise a fresh modified search based on the present ASK of the user is initiated.
- The process continues till such time the ASK of the user is resolved.

**14.3.1** The design of a cognitive user model presents several issues that need to be considered. As mentioned earlier, this is a recent development in IR and as such only a few experimental systems and studies are available. The most important of the requirements for a cognitive IR model is an effective mechanism for user-computer interaction / dialogue. The designer should have a clear idea of the nature of interaction that should take place between the user and the system and also the kind of user model that the system should build to make the dialogue meaningful and beneficial in retrieval

terms. An early model was developed by Oddy and was called **THOMAS\***. A description of this model is given below to provide an idea of the difference in approach between cognitive user modeling and other models of information retrieval.

The THOMAS system developed by Oddy was characterized by the fact that it did not require a user to input a query. This was based on the fact that many end users find it difficult to submit a well-formulated query at the beginning of a search. Instead the user was required to enter titles of documents, authors of interest and / or keywords. Starting from this initial input the system was designed to lead the user through interactions, to identification of interesting (*'relevant'*) items. In actual implementation of the system the process employed by THOMAS was broadly as below:

- Go to a portion of the knowledge base\* stored in the system that closely matched the initial inputs of the end user in terms of textual similarity
- Display the documents associated with the terms
- Prompt the user to:
  - Suggest for each document displayed, whether the document is of interest or not and / or
  - Select or reject terms in the representation (metadata) of a document
  - Suggest additional keywords or documents
- Use these additional inputs to refine the system's understanding of the user's requirements by modifying the original profile created in the beginning of the search and create a better user model; In fact refining the model of the user based on the world model of the knowledge base is a something that continuously happens in THOMAS.

---

\* Another experimental system based on cognitive user modeling is GRUNDY developed by E. A. Rich though this system adopted a different approach to building the user model. GRUNDY experimented with fiction retrieval

• The program's knowledge base was essentially a network with the nodes consisting of terms, authors, and documents linked to one another indicating associations

It is important to have a general idea of the way the database was structured to function in a manner described above. It was built as a network of associations involving three elements, viz., documents, authors and subject terms. The association could be between any combinations of these elements. During the process of interaction starting from the initial input by the user, a model of the user is built which is continuously updated. At any particular point of time during the information retrieval process the user model is represented as a 'context graph'. Again, the context graph basically consisted of nodes (involving the same three elements as above). Any document in the database which had a 'high degree of involvement' with the user model was presented to the user for his / her response during the process of retrieval / interaction.

**14.3.2** Cognitive models approach information retrieval as a problem of developing an adequate model of the user's requirements. The user model created by the system is used to identify documents for retrieval from the system's database. A major limiting factor in such an approach could be the methodology employed for building the term association network involving keywords. In order to be effective the association between terms should be based on semantics. Whether associations based purely on terms will prove adequate and effective in large-scale systems is a question.

Another important limiting factor is that which could be imposed by the nature of interactions that actually take place between the system and the user during the search process. In order to be effective it is necessary that the communication between the user and the system does not break down and is along the desirable lines. In other words there are limitations to the ability of a program to build complex models of human cognition primarily on the basis of inputs made during a dialog with the system.

**Self-Check Exercise**

**2. Write a brief note on Cognitive Model of IR?**

**Note:**

**i). Write your answer in the space given below.**

**ii). Check your answer with the answers given at the end of this Unit.**

.....

**Self-Check Exercise**

### **3. What are the steps involved in the process of Information Retrieval in**

**Cognitive point of view?**

**Note:**

**i). Write your answer in the space given below.**

**ii). Check your answer with the answers given at the end of this Unit.**

.....

#### **14.4 CANSEARCH**

CANSEARCH was another interesting example of an expert intermediary system designed to help users searching the MEDLINE database. The system is important in that it adopted a different approach to assisting the user. CANSEARCH was intended to assist novice users in building an appropriate search formulation for searching documents on cancer included in the MEDLINE database. CANSEARCH helped users to construct an appropriate search by taking them through relevant MeSH terms. It employed as its knowledge base basic knowledge of cancer treatment, MeSH terms for the domain, and knowledge of rules related to the indexing procedure employed by MEDLINE. CANSEARCH was not designed to carry out a search in the bibliographic database but was limited to help the users formulate a complete search strategy appropriate to their needs. The search strategy formulated with the help of CANSEARCH could then be employed in a search of the MEDLINE database to retrieve relevant bibliographic records.

#### **14.5 PLEXUS**

PLEXUS was another domain-specific expert system (covering the domain of Gardening). PLEXUS is of particular interest in that it made use of a hierarchical classification based on the relevant schedules from the Broad System of Ordering (BSO) of FID. The terms were categorized using a schema of conceptual categories. It also made use of a dictionary. All these were part of the knowledge base of the system. Beginning with a request to the user to provide a problem description, PLEXUS built a model of the user based on the input containing the initial problem description and subsequent interactions.

•PLEXUS had a procedure for dealing with words that did not match with any term in the dictionary. If PLEXUS did not recognize any of the terms in the initial input by the user, the user was asked to browse the hierarchical BSO display and select appropriate terms.

A rough description of the procedure adopted by PLEXUS is given below:

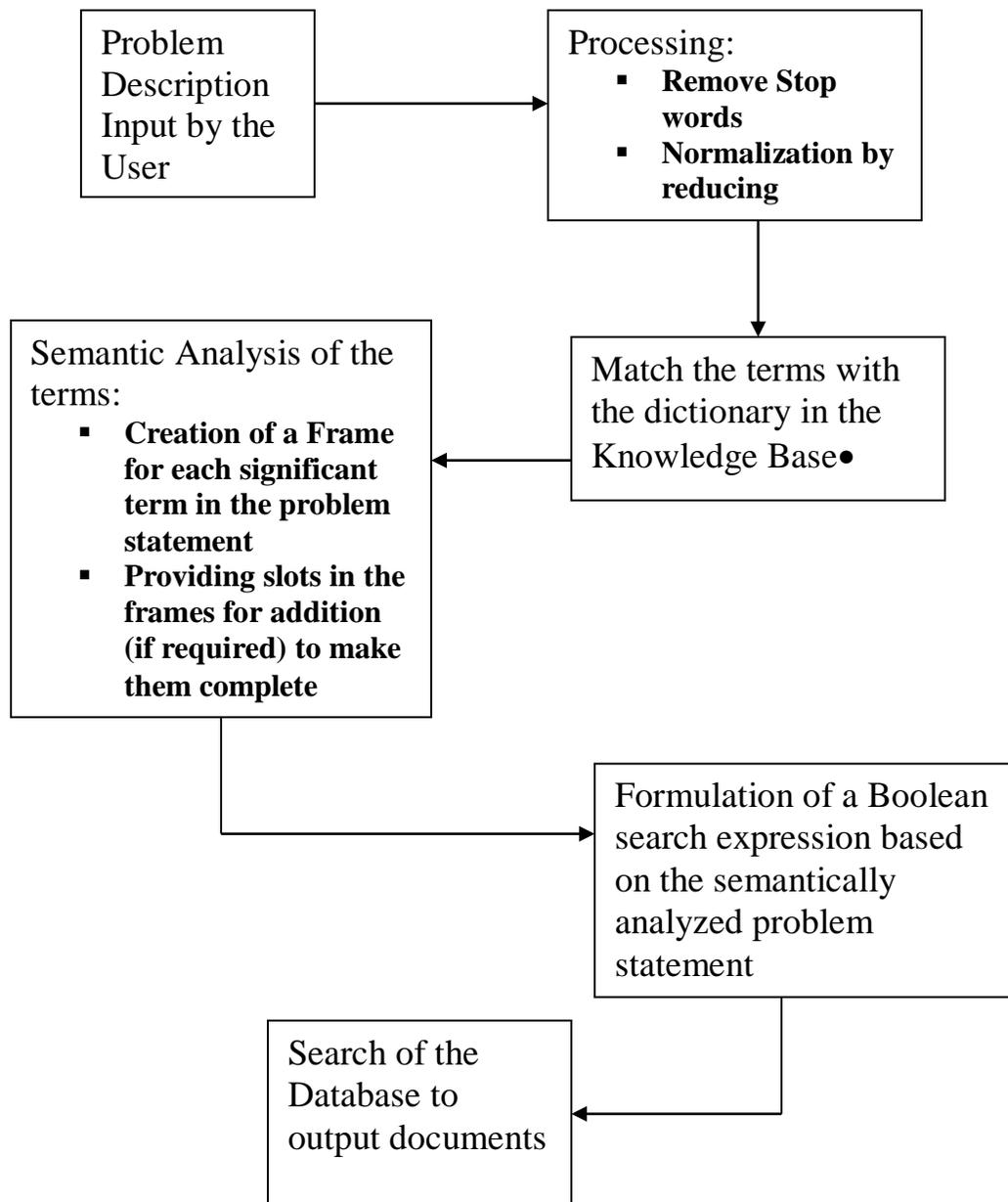


Figure : The PLEXUS System

## **14.6 I<sup>3</sup>R (*Intelligent Intermediary for Information Retrieval*)**

There have been other experiments such as I<sup>3</sup>R (*Intelligent Intermediary for Information Retrieval*), which employed expert system technology for information retrieval and many others. In general expert systems' approach to information retrieval is centered on the idea that in order to optimize retrieval performance, information retrieval systems should have an adequate and accurate picture of the user's information requirements. Starting with this basic premise expert intermediary systems attempt to improve the performance of IR systems by seeking to enhance the user's query using domain knowledge. The domain knowledge is generally obtained either from the user or from domain-specific knowledge tools such as a micro-thesaurus, classification systems, etc. The other focus in expert system-based information retrieval has been on enhancing the interaction between the user and the system.

## **14.7 Hypertext/ association approach to IR**

Hypertext links provide a means of directly connecting two distinct pieces of text. The concept was developed in the 1960s and enjoyed a certain vogue, particularly among educators who saw it as a means of developing better educational programs. The idea languished for about two decades but has come very much alive with the advent of the World Wide Web. Virtually every web page contains several hypertext links to other pages, often in parts of the Web developed by somebody else. A hypertext link consists of an identifier and a pointer. The identifier is often a highlighted word or phrase in the text but can be an explicitly given name. In books using a hypertext style, the pointer is often a page number. In this sense, a hypertext link is much like a "see also" reference. In the Web, the hypertext link is active: when the user clicks on the identifier, a transfer to the linked page occurs.

## **14.8 Hypertext Model**

The term '*Hypertext*' was coined by Ted Nelson. Literally *hypertext* means additional dimensions to texts. By usage it refers to a facility that allows users of texts to jump from one block of text to another block of the same text or some other text. Conceptually however, the notion of hypertext is quite similar to the idea of cross-references employed by Cutter and other cataloguers to link semantically related subjects. An early project in the area, called HYPERCATALOG\* As a technology it is extensively used to build and use associated texts by navigating between associated texts that are linked using hyper-linking technology. Hypertext systems can be visualized as blocks of texts that are networked in a seamless fashion to facilitate navigation between related blocks of texts. In reality hypertexts allow linking of and navigation between different kinds of digital data (text, video, audio, multimedia, etc); the term Hypermedia is used when it comes to creation of hyperlinks between discrete units of different kinds of data – text, graphics, audio, video, etc.

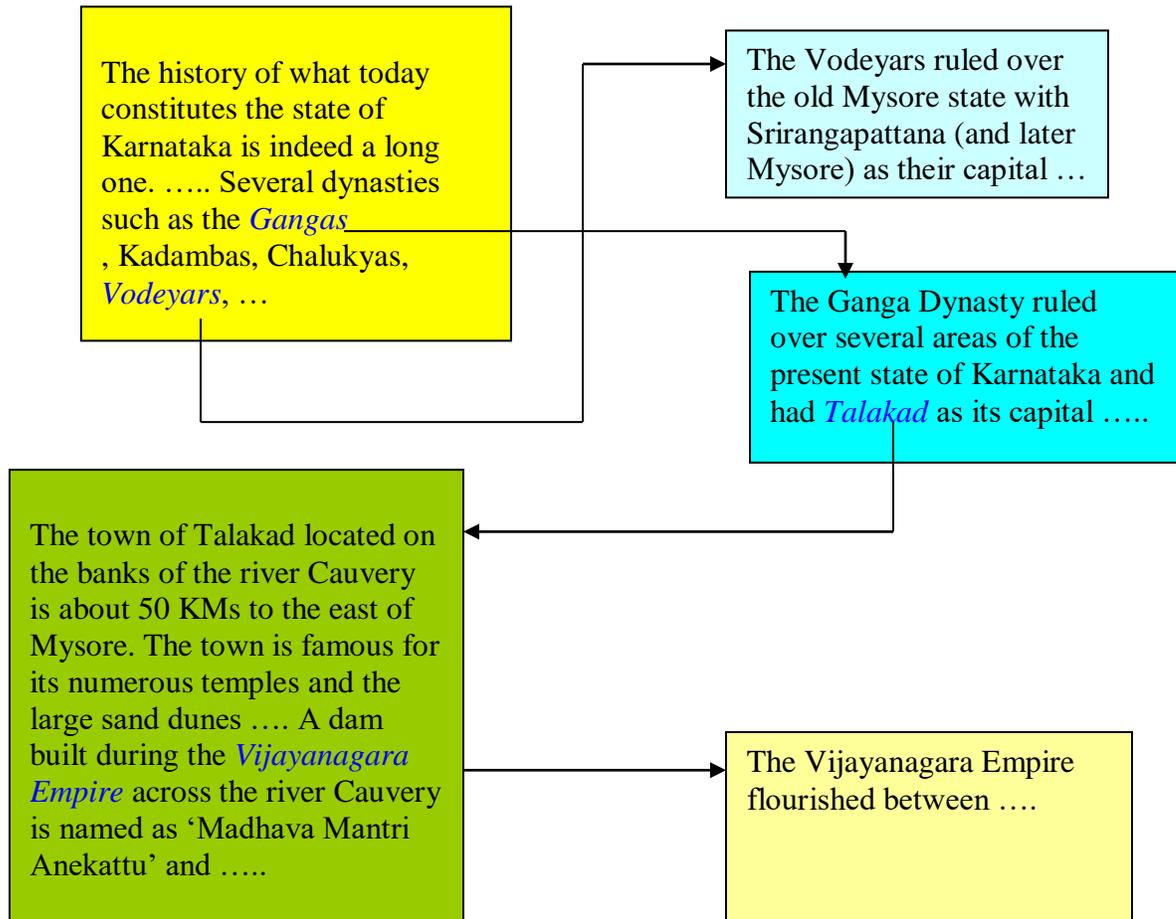
Hyper-linking of blocks of documents creates an interactive navigational structure that allows the end users to browse related documents / data / information in a non-sequential fashion. In effect the blocks of texts (or other forms of data) can be considered as nodes, which are connected using hyperlinks. The linked nodes are semantically or otherwise, related to one another. In fact the World Wide Web (WWW) is based on this technology. The normal convention that is employed in hypertexts is to attaché the hyperlink to a string of text (or to an Icon). Such strings / icons are specially marked to indicate that they are hyperlinks (usually the string is underlined or appears in a different font colour and so on). In the course of reading a text the user might come across such marked strings / icons. If the user clicks on the marked string / icon the directed link is followed and the user is taken to the linked document / portion of the document.

Hypertexts are used for navigation / browsing to find some required information. For example, a student of History of Karnataka may be interested in finding more information on the *Ganga* Dynasty or about Talakad (Which was the capital of the *Gangas*). A general document on the History of Karnataka could be linked to a more detailed text on

---

\* The **HYPERCATALOG** was an international venture involving the University of Linkoping, Sweden, the University of Tampere, Finland and Informatics Management and Engineering, England; The term 'HYPERCATALOG' to name the project was employed to indicate both the connection with Hypertext and also that it was an enhancement and extension of the traditional catalogues.

the Ganga Dynasty and / or to Talakad allowing the user to jump to these resources if required.



**Figure 7: Hypertext Model**

*Note: The coloured fonts are hyperlinks and clicking on these will take the user to the document / portion linked.*

Hypertext systems are often perceived as systems that facilitate browsing for relevant information. However since browsing is often used in retrieving relevant information, Hypertext is also considered a retrieval model. There are fundamental differences between **Hypertext model and the other models of information retrieval and also certain important issues associated with hypertext model.**

- In practice when a hypertext is very large, the user may lose track and navigate to portions / texts that are not directly related to his requirements. This can be

avoided if the hypertext adopts a map showing the path the user navigated to arrive at a particular node.

- The user is restricted to the flow of information conceived and formulated by the designer of the hypertext. It is important for the designer of the hypertext to take into consideration the needs of the potential users; otherwise it is highly probable that the hypertext may not provide all the required links that a potential user may want to follow. This necessitates an understanding of the user needs through a requirement analysis carried out before the hypertext is designed.
- Even with a map, navigation in hypertext could still prove to be difficult and less productive because of excessive hyper-linking at the time design of the hypertext. A model that is beneficial in this context is to adopt a more structured approach to the design of the hypertext. For example the hypertext can be organized hierarchically to facilitate easy navigation.
- Another possibility is to attach weights to links so as to enable users to identify the more important links to follow. In order to be meaningful and useful this weight has to be a function of the path that the user is coming from and the same link will have different weights depending on the path of the user.
- IR systems traditionally carry out a direct search of the database using as input a query from the user. As against this in the hypertext model, search is carried out by a process of navigation. The user has the flexibility to build dynamically an information path; however, this could be time-consuming and could also lead the user away from his initial requirement
- Unlike the conventional IR systems in which the process of information retrieval gets initiated with a user query, the Hypertext model is a browsing model and reflects a situation similar to what normally obtains in a small library when the user is quite familiar with the collection and location of resources on his areas of interest. In such a situation the user normally goes directly to the section containing resources of his interest

and browses to identify any new resource or a familiar resource. But in large collections or unfamiliar resource collections the user normally begins with a query (even if the query is fuzzy). In large hypertext systems it is useful to assist the user in finding the area to begin the browsing and this can be done only by querying the system which will take the user to that section of the hypertext that will probably be the best starting point to begin browsing.

Hypertext and hypermedia are extensively used and can be said to have become an important method of information retrieval since the arrival of the World Wide Web. A major area of application of the hypertext model in information retrieval is in linking source documents to cited items and to their full texts. The principal advantage of hypertext in this context is the ease with which references can be followed. Interest in hypertext for a variety of applications has been growing since the 1980s.

### **Self-Check Exercises**

#### **1. What are the issues associated with the Hypertext IR Model?**

**Note:**

**i). Write your answer in the space given below.**

**ii). Check your answer with the answers given at the end of this Unit.**

.....  
.....

### **14.9 Expert Systems**

Cognitive user modeling has to necessarily employ a certain amount of AI in order to be able to build adequate user models. It is therefore not surprising that an area of research interest in information retrieval has been that related to building effective expert intermediary systems that will help carry the interactive process between a user and the system in a meaningful and useful manner. Both expert intermediary systems and cognitive user models share something in common in that they try to automate certain aspects of information retrieval. Expert intermediary systems in IR generally have the

aim of functioning in the place of a human search intermediary who would normally have a dialogue with the user in an effort to ascertain his / her requirements and develop a search strategy and build a search. Naturally therefore, there has been focus on simulating the nature of interaction that normally takes place between a user and a human search intermediary. In other words modeling the interaction between the user and the search intermediary has been a major focus of research in this area. To a certain extent the objective has been to develop an expert intermediary system that would serve as a front-end. There are two broad approaches in using expert intermediary systems in the context of information retrieval. In some experiments, the function of the expert system is limited to assisting the user to formulate an appropriate search. The database is not a part of the system. As such the expert intermediary system by itself could not be employed to carry out the actual search of the database to retrieve relevant references. Such systems were designed to accept user terms and transform them into command statements using the search language. The search formulated by the expert system was then used in an online database and the actual search carried out. Obviously such intermediary systems are designed to work with specified online databases and are generally designed to work with a known domain-specific database. There have also been experiments at developing more complete information retrieval systems that employ an expert intermediary system as the front-end. In such systems the expert intermediary system assists the user (just as a human intermediary would) to formulate an appropriate search strategy before carrying out the actual search in the database, which is an integral part of the system.

Just as in cognitive user modeling, expert intermediary systems recognize that the principal issue in information retrieval is related to the problems users have in unambiguously expressing their information requirements. This is seen as a pre-requisite for effective information retrieval. Expert intermediary systems have experimented with a number of different approaches to address this issue.

An early system developed was CONIT, which considered the problem users face in learning and understanding different search languages used by different database hosts. CONIT was built to help access three different database hosts, viz., NLM (MEDLINE), SDC and DIALOG. In effect CONIT would translate the user requests to the command

languages of the system and carry out searches. The responses of the system were also translated to a common language before presenting the results (output) to the user. CONIT gives the searcher the feeling that he / she is interacting with a single system rather than multiple systems.

An issue that has been debated in the literature on the subject is the applicability and appropriateness of expert systems for information retrieval. Many are of the opinion that information retrieval is not an ideal problem area for application of expert system technology. The lack of predictability and homogeneity in the process of information retrieval are the major factors making it difficult to develop expert information retrieval systems. Nevertheless, it must be acknowledged that expert systems have had significant impact on research in information retrieval and have influenced a shift in the focus from retrieval algorithms to a paradigm concerned with users, domain knowledge and human-computer interaction.

#### **14.10 Summary**

Two algebraic IR models are the latent semantic indexing and the neural network models. The Latent Semantic Indexing proceeds on the premise that index terms derived from texts of documents often lead to poor retrieval performance in view of the limitations of the related index term derivation processes in summarizing the contents of documents adequately and accurately. The Neural Network Model seeks to adopt the approach of pattern matching used by human brain in the IR process. A number of limitations of natural language text in serving effectively as the basis for deriving index terms have been discussed in related literature. The origins of the cognitive model of IR can be seen in the '*Anomalous state of knowledge*' (ASK) hypothesis of Belkin. It is therefore useful to have a broad understanding of the rationale behind ASK. The ASK notion can be discussed and explained only in a cognitive framework.

#### **14.11 Answer for the Self Check Exercises**

##### **1. What are Algebraic Models of IR?**

Two other algebraic IR models are the **latent semantic indexing** and the **neural network** models. The Latent Semantic Indexing proceeds on the premise that index terms derived from texts of documents often lead to poor retrieval performance in view of the limitations of the related index term derivation processes in summarizing the contents of documents adequately and accurately

The **Neural Network Model** seeks to adopt the approach of pattern matching used by human brain in the IR process. In other words a neural network IR model is a simplified version of the mesh of interconnected neurons in the human brain.

## **2. Write a brief note on Cognitive Model of IR?**

The origins of the cognitive model of IR can be seen in the '*Anomalous state of knowledge*' (ASK) hypothesis of Belkin. It is therefore useful to have a broad understanding of the rationale behind ASK. The ASK notion can be discussed and explained only in a cognitive framework. Let us imagine a scholarly communication situation. When an author writes a paper or some other document containing scholarly information, the text of the document is essentially a transformed representation of the cognitive knowledge structure of the individual author. When another individual studies / reads this document his cognitive structure interacts with the cognitive structure of the author of the document resulting in a transformation of the cognitive knowledge structure of recipient of the communication. The cognitive knowledge structure of an individual in respect of an entity at any particular point of time is his / her worldview of that entity. This is not static and keeps on changing as the individual receives communication related to the entity. . The most important of the requirements for a cognitive IR model is an effective mechanism for user-computer interaction / dialogue.

## **3. What are the steps involved in the process of Information Retrieval in Cognitive point of view?**

The process of information retrieval involves the steps when viewed from a cognitive point of view, A user initiates a search in a system recognizing an anomaly in his cognitive structure in respect of an entity (subject), This is converted into and expressed

in the form of a request (query) and submitted to a retrieval system, The retrieved texts are examined by the user which in essence means that the cognitive structure of the communicator (author) of a retrieved text interacts with the cognitive structure of the user, If the recipient feels that the anomaly in his / her knowledge structure has been resolved, the process is brought to a stop; otherwise a fresh modified search based on the present ASK of the user is initiated, The process continues till such time the ASK of the user is resolved.

#### **14.12 ANSWERS TO SELF-CHECK EXERCISE**

##### **1. What are the issues associated with the Hypertext IR Model?**

In practice when a hypertext is very large, the user may lose track and navigate to portions / texts that are not directly related to his requirements. The user is restricted to the flow of information conceived and formulated by the designer of the hypertext. navigation in hypertext could still prove to be difficult and less productive because of excessive hyper-linking at the time design of the hypertext. Another possibility is to attach weights to links so as to enable users to identify the more important links to follow. Search is carried out by a process of navigation. The user has the flexibility to build dynamically an information path; however, this could be time-consuming and could also lead the user away from his initial requirement

#### **14.13 Key Words**

**Expert System:** An inferential information system built on a knowledge base.

**Neural Networks:** An algebraic model of document retrieval based on representing query, index terms, and documents as a neural network.

**Intelligence Information Retrieval:** It is a computer system with inferential capabilities such that it can use prior knowledge to establish a connection between a user's request and a candidate set of relevant documents.

Considering the fact that Information Retrieval as a sub-discipline of Information Studies has a relatively short history compared with some other sub-disciplines, it does indeed

appear that there is a vast amount of discussion on Information retrieval. This is because of the centrality of information retrieval to the discipline of information studies. The focus and emphasis have changed over the decades; beginning with the Aslib-Cranfield studies, which focused on improving indexing languages as the primary means to enhancing retrieval effectiveness to focusing on understanding the user's requirements and user-system interaction in the cognitive user models. The very nature of information retrieval and the entities involved in the process of IR make it difficult to develop a universally acceptable model of IR. There are several reasons for this. First of all *information* is a cognitive and not a physical entity and the notion of *Relevance* is highly subjective. The problem of measurement has, therefore, remained. These probably explain why progress in information retrieval has been slow.

**Hypermedia:** It refers to the creation and representation of links between discrete pieces of different kinds of data- text, numbers, graphics, and sound.

**Hypertext:** It refers to a facility that allows users of texts to jump from one block of text to another block of the same text or some other text. It is system to manage a collection of information that can be accessed non-sequentially.

**Expert System:** An inferential information system built on a knowledge base

#### 14.14 Recommended Readings:

1. Chowdhury, G. G. Introduction to modern information retrieval. – London: Library Association Publishing, 1999 (especially chapters 8, 16 and 17)
2. Ellis, David. Progress and problems in information retrieval. London: Library Association Publishing, 1996
3. Baeza-Yates, Ricardo and Ribeiro-Neto, Berthier. Modern information retrieval. – Delhi: Pearson Education, 2004
4. Baeza-Yates, Ricardo and Ribeiro-Neto, Berthier. Modern information retrieval. – Delhi: Pearson Education, 2004

5. Chowdhury, G. G. Introduction to modern information retrieval. – London: Library Association Publishing, 1999
6. Ellis, David. Progress and problems in information retrieval. London: Library Association Publishing, 1996

---

## **Unit -15**

### **Retrieval effectiveness**

---

#### **Structure**

15.0 OBJECTIVES

15.1 INTRODUCTION

15.2 SYSTEM CENTRED MEASURES AND USER CENTRED MEASURES

15.3 EVALUATION MEASURES AT TREC CONFERENCE

15.4 RETRIEVAL EFFECTIVENESS MEASURES

15.5 EVALUATION OF RETRIEVAL EFFECTIVENESS

15.6 EVALUATION VIEWPOINTS AND THE RELEVANCE PROBLEM

15.7 TRENDS AND ISSUES

15.8 EFFICIENCY MEASURE

15.9 RELEVANCE

15.10 RECALL

15.11 PRECISION

15.12 RECALL AND PRECISION DEVICES

15.13 FACTORS AFFECTING RECALL AND PRECISION

15.14 RETRIEVAL PERFORMANCE EVALUATION

15.15 EVALUATION OF SYSTEM COST AND EFFICIENCY

15.16 SUMMARY

15.17 ANSWERS TO SELF-CHECK EXERCISES

15.18 GLOSSARY OF TERMS

15.19 RECOMMENDED BOOKS FOR FURTHER READING

## **15.0. OBJECTIVES**

- To describes the origins of IR research in the manual testing of indexing languages and indexing language devices
- To present an overview of the system centred measures and user centred measures in retrieval research
- To help you interpret the significance of the results

### **15. 1. Introduction**

The purpose of organization of information is timely access. There have been several innovations in the past on the long term storage aspects but little has been done on information access. Information retrieval (IR) focuses on the study of process involved in the storage of access of to and representation of information continuers along with the system that implement IR concepts (wolfram). IR has attracted a growing number of investigators as efficient storage and effective access to increasingly proliferating information must be managed in many environments. Today most of the IR is computers based. In this unit the IR is used to describe predominantly, document retrieval systems.

Wolfram, Diet (2005), Applied informetrics for information retrieval research. A West port library unlimited Inc., Information retrieval research is a diverse field and it is impossible to provide any overall conclusion of the state of the art.

System centred measures evaluate the efficiency and effectiveness of an IR systems performance. The effectiveness of an IR system, or how well it works, depends not only on its ability to retrieve relevant documents, but also its ability to exclude non-relevant documents. These are not simply two sides of the same coin. Systems may be quite effective in retrieving relevant documents, but in doing so, additional non-retrieval documents may be retrieved and must be filtered out by the user.

## 15.2 Origins of retrieval research

USA the origins of information retrieval research can be traced back to 1953. Then separate tests were carried out in Britain and the United States evaluating the performance of the, at the time, controversial Uniterm system devised by Mortimer Taube. Which represented documents by single terms taken from titles or abstracts, against more conventional approaches to subject indexing and retrieval? These two tests were the Armed Services Technical Information Agency (ASTIA) —Uniterm test carried out in the United States, which was reported by Gull, and the Cranfield-Uniterm test undertaken at the College of Aeronautics. Cranfield, in the United Kingdom and described by Thorne.

In the ASTIA — Uniterm test, two groups — one consisting of the indexing staff of ASTIA, the other of staff of Mortimer Taube's company, Documentation Incorporated - separately indexed and then searched the ASTIA collection, which then consisted of around 15,000 documents, with 93 requests which had been submitted to ASTIA in the normal course of its activities. The ASTIA staff indexed the documents, employing the operational ASTIA alphabetical subject headings list; the Documentation Incorporated staff used uniterms. The measure of effectiveness employed by the two groups was that of relevance of documents to the question. This appears to be the first appearance of relevance as a performance criterion for information retrieval system evaluation and, in this respect, at least in the case of the USA, information retrieval system testing and the employment of relevance as a performance criterion made their entrance together.

## **Experience in UK**

In the same year as the ASTIA - Uniterm test in the USA, a completely separate test of the performance of the Uniterm system against more conventional forms of indexing was carried out in the UK at the College of Aeronautics, Cranfield, Bedfordshire. This test has been described by Thorne. In its objectives, the Cranfield-Uniterm test was very similar to the ASTIA-Uniterm test. Again, the comparison was of the performance of a traditionally-based operational indexing system - in this case a modified version of the Universal Decimal Classification (UDC) in use at the Royal Aircraft Establishment (RAE) - and that of the Uniterm system.

The Cranfield-Uniterm test adopted a very different approach to that followed in the ASTIA — Uniterm test. In the first place, rather than using the actual document collection, a limited collection of 200 documents on the subject of aeronautics was set up. A selection of documents from this collection, referred to as the source documents, was then employed to derive 40 artificial requests, to which, it was argued, individual source documents represented answers. The collection of 200 documents was then searched with the 40 artificial requests. The criterion of effectiveness was that of success in retrieving the source documents - that is, the documents from which the requests employed had been originally derived. The procedure represents an attempt to side-step the problem of deciding which documents are relevant to which requests by making the requests relevant to the source documents. The test results demonstrated a relatively outstanding performance by the Uniterm system: 85% of the source documents were retrieved using uniterms, compared with only 50% employing the UDC/RAE classification

## Testing indexing systems: Cranfield I

The grant, which was awarded in July 1957, was for an investigation of the comparative performance of four indexing systems. The systems to be tested were UDC, an alphabetical subject index, a faceted classification scheme and the Uniterm system.

The Cranfield I test was roughly similar in conception and execution to the preceding Cranfield-Uniterm test, although on a much more ambitious scale. In this case a collection of 18,000 documents on aeronautical engineering, about half of which dealt with the subject of high speed aerodynamics, was indexed using each of the four indexing systems to be tested. A set of 1,200 questions was elicited, based again on the source document principle. The collection was then searched with this set of questions. If the source document was identified, then the search was deemed successful; if not, it was considered to have failed. Searches which failed to identify the source document were subsequently analysed to discover whether the failure to identify the source document was the result of difficulties with the question, the indexing, the searching or the system

### **Level of Performance**

The results of the test seemed to show that all of the systems tested operated at broadly the same level of performance, at least in terms of the ability to retrieve the source documents. In summary, the results were that source documents were retrieved in the following percentage of all the searches carried out for each of the systems tested:

Uniterm	82.0%
Alphabetical subject headings	81.5%
UDC	75.6%
Facet classification scheme	73.8%

Once again, the results indicated that the Uniterm system performed well as. There was the added advantage that the Uniterm system required no intellectual effort at the indexing stage.

**Supplementary Tests:**

First, a sub-set of 100 of the test questions was sent out to library and information workers in other organizations in appropriate fields. These individuals were asked to compile a list of references based on the questions sent. In the event, 88 lists were prepared and returned, and the contents of these were then checked for their availability in the original document collection tested. Of these 8 had no references in common with those in the collection, and the remaining 80 provided, in all, 359 different references. Each of these 359 references was then assessed for relevance in relation to the source document and the appropriate question. Three levels of relevance were considered: as useful as the source document, somewhat useful and not useful. 'At the first level, it was considered that there were 53 documents which related to 35 questions; at the second level there were 67 documents which related to 6 questions. The 41 questions were then used to search the entire document collection. The success rate in identifying these documents was somewhat different from that achieved when using the source documents alone; notably, the performance of Uniterm and alphabetical subject headings was identical, and that of UDC nearly so:

Uniterm	75.0%
Alphabetical subject headings	75.0%
UDC	74.0%
Facet classification scheme	60.0%

In the second of the supplementary tests, a random sample of 79 of ' the searches carried out for the original requests was analysed to identify the total number of documents actually retrieved in the course of each of those searches. A sample of 759 of the retrieved documents was then assessed for relevance using the same method as that employed in the other supplementary test. This was undertaken in an attempt to provide some indication of the effectiveness of the different systems in terms of their performance in holding back non-relevant material; that is, on the extent to which searchers were retrieving only relevant material. Again the Uniterm system performed well, although this time not quite as well as the alphabetical subject headings system:

Alphabetical subject headings	12.5%
Uniterm	12.0%
Facet classification scheme	7.5%
UDC	7.0%

However, again the figures need to be treated with caution. It was found possible to derive entirely different figures, particularly if the searches had been carried out beyond the point of retrieval of the source documents.

From these two supplementary tests, Cyril Cleverdon, the Cranfield I project director, attempted to rebut the two criticisms of the main testing procedure outlined above. Cleverdon argued that the performance of the systems tested, in terms of the retrieval of relevant non-source documents, was roughly similar to that of the retrieval of source documents, given identical source document-based requests. Cleverdon also attempted to derive figures for the retrieval of non-relevant documents, and contended

that the tests demonstrated that there was an inverse relationship between the ability to retrieve relevant documents and the ability to exclude non-relevant ones. This relationship, which Cleverdon originally contended had been conclusively demonstrated in the Cranfield 1 tests, was subsequently presented as a hypothesis in the Cranfield II tests.

Other tests carried out in conjunction with the Cranfield institute, at around the same time as the Cranfield I test, were of the English Electric Library faceted classification at Whetstone, UK, and of the Western Reserve University (WRU) index of metallurgical literature in the USA. It is not intended to go into detail into these tests here; the results of the tests were broadly in line with those of Cranfield 1, as were the methods employed. However, two features of the Cranfield -WRU test are of particular interest:

- (a) The efforts made to obtain exhaustive relevance assessments for all the documents in the collection.
- (b) The attempt made to calculate reliable figures for 'recall' and 'relevance'. Recall referred to the proportion of those documents in the collection which was held to be relevant and which was actually retrieved; 'relevance' to the proportion of those documents retrieved which was held to be relevant.

In later writings the second ratio was referred to as the precision ratio to avoid confusion with the more general concept of relevance on which both the recall and the precision ratios are based.

The Cranfield I test was subjected to extensive criticism. The employment of source documents both to derive the queries and to provide the basis for evaluating retrieval effectiveness was particularly singled out for criticism. Swanson considered that

most of the results of the tests were due to the use of source documents in this dual role. There were two aspects of this criticism - that in an operational situation a source document generally does not exist, and that the relationship between source document and query was too close.

Swanson noted that there was a close verbal similarity between the query terms and the terms used in the source document, so much so that simply matching query terms with source document title terms would have provided a recall figure of 85%. If this were the case, and there were a tendency for the terms used in the source documents actually to influence the terms used in the queries, then the Cranfield I test could be said to be inherently biased in favour of a term-based system like Uniterm and against concept-based systems such as classification schemes. In order to remove this bias, or potential bias, Swanson argued that the source documents should have been excluded from the tests.

The consensus of opinion was that although they represented a significant step forward in the evaluation of indexing systems, the Cranfield I and associated tests were seriously methodologically flawed by the employment of source documents for evaluating retrieval effectiveness. It was felt that any subsequent tests should not employ source documents but instead should base performance evaluation on the retrieval of any relevant documents. Sharp, in a review of the Cranfield-WRU test, put the point succinctly: 'the source document principle should be dropped and future tests carried out taking into account all relevant documents retrieved (p. 174)

## **Testing indexing devices: Cranfield II**

The second major series of tests undertaken at the Cranfield institute was made possible by a further grant to Aslib (UK) from the National Science Foundation (USA). The Cranfield II tests were rather different in form from the previous tests carried out in that, rather than being tests of existing operational systems or models of such systems, the tests were of different indexing devices. Indexing languages, it was considered, consist of amalgams of such devices, and so for the Cranfield II tests 33 different types of indexing languages were constructed with varying terminologies and structures. The different indexing languages varied in the extent to which they used single or compound terms and hierarchies, and were controlled for synonyms or homographs.

The Cranfield II testing environment was compared by Cleverdon to that of a wind tunnel, the intention being to control any variable not being tested and to obtain results which/could be used to make statements about retrieval systems in general. The document base consisted of material on aeronautics; in this case the document sample consisted of 1,400 items. A total of 211 requests was obtained by asking authors of selected published papers (the base documents) to reconstruct the questions which originally gave rise to the writing of these papers. In order to speed up the conduct of the tests, most of the searches were carried out employing smaller sub-sets of documents and queries, the smallest consisting of 200 documents and 42 requests. Searching was carried out using search strategies based on different levels of coordination of terms

A significant difference between the testing procedure employed in the Cranfield II tests and that of the earlier Cranfield I test was that the measure of effectiveness was now explicitly relevance-based. Performance was judged by the retrieval of items

previously identified as relevant to the question. The relevance of documents in the collection to the search questions was determined prior to any searches being carried out by the employment of a two-stage procedure. In the first stage, students of aeronautics searched the entire document collection. The documents which they identified as relevant, plus the references contained in the base document, were then sent to the author/requester for final relevance judgement. In addition, a number of references were also identified which had citations in common (bibliographic coupling), and these too were sent for final relevance judgement by the author/requester. The tests were carried out to investigate the effect on retrieval performance of varying the different generic index language features and, more specifically, to examine the effects on retrieval performance of devices intended to increase precision and those intended to increase recall. Recall was calculated by dividing the number of documents in the collection held to be relevant to a question by the number of those documents actually retrieved:

$$\text{Recall} = \frac{\text{Relevant documents retrieved}}{\text{Relevant documents in collection}}$$

Precision was calculated by dividing the number of relevant documents retrieved by the number of documents retrieved:

$$\textbf{Precision} = \frac{\text{Relevant documents retrieved}}{\text{Documents retrieved}}$$

The relationship between recall and precision was obtained for different coordination levels for individual searches and average figures were derived by normalizing the results

of a number of different searches. The findings of the Cranfield II tests were, in summary, that:

- (a) The best performance was obtained by using single-term index languages.
- (b) When employing single-term index languages, the formation of groups of terms or classes beyond true synonyms or word forms resulted in a drop in performance.
- (c) Use of devices to increase precision other than that of class coordination is not as effective as simple coordination.
- (d) Natural language with confounding of synonyms and word forms combined with simple coordination provided a reasonable performance.
- (e) Every set of figures supported the hypothesis that there was an inverse relationship between precision and recall.

The results of the Cranfield II tests - particularly the conclusions relating to the superiority of single-term natural language indexing and to the existence of an inverse relationship between precision and recall - were widely reported and debated. The findings were contrary to almost all expectations of the library and information science community. That single terms taken from natural language, with minimal control for true synonyms or word forms, combined with simple coordination would outperform indexing systems which had been the subject of the most careful design and creation astonished researchers and practitioners alike. The existence of an inverse relationship between recall and precision carried the further implication that attempts to maximize both in a single system were bound to fail Relevance as a performance criterion

### *Relevance as a performance criterion*

Cleverdon had correctly seen that if relevance was to be employed as a performance criterion, the form of relevance judgement employed had to be objective enough to serve as the basis of the measure of effective-ness. However, the problems experienced with simulating relevance judgements in the course of the Cranfield tests, and the studies undertaken by Cuadra and associates and by Lesk and Salton, demonstrate that it is difficult to simulate real-life relevance judgements in a way which will both be reasonably authentic and provide confidence that the simulated judgements will be consistent and valid

Before the final implementation of an information retrieval system, an evaluation of the system is usually carried out. The type of evaluation to be considered depends on the objectives of the retrieval system. Clearly, any software system has to provide the functionality it was conceived for. Thus, the first type of evaluation which should be considered is functionalities. These are tested one by one. Such an analysis should also include an error analysis phase in which, instead of looking for functionalities, one behaves erratically trying to make the system fail. It is a simple procedure which can be quite useful catching programming errors. Given that the system has passed the functional analysis phase, one should proceed to evaluate the performance of the system.

The most common measures of system performance are time and space. The shorter the response time, the smaller the space used, the better the system is considered to be. There is an inherent tradeoff between space complexity and time complexity, which frequently allows trading one for the other.

In a system designed for providing data retrieval , the time and the space required are usually the metrics of most interest and the ones normally adopted for evaluation the system. In this case, we look for the performance of the indexing structures, the interaction with the operating system, the delays in communication channerls, and the overheads introduced by the many software layers which are usually present. We refere to such a form of evaluation simply as performance evaluation.

In a system designed for providing information retrieval, other materics besides time and space, are also of interest. In fact, since the user query request is inherently vague, the retrieved documents are not exact answers and have to be ranked according to their relevance to the query. Such relevance ranking introduces a component which is not present in data retrieval systems and which plays a central role in information retrieval. Thus, information retrieval systems require the evaluation of how precise is the answer set. This type of evaluation is referred to as retrieval performance evaluation.

## **15. 2. RECALL AND PRECISION**

Consider an example information request  $I$  (of a test reference collection) and its set  $R$  of relevant documents. Let  $R$  be the number of documents in this set. Assume that a given retrieval strategy processes the information request  $I$  and generates a document answer set  $A$ . Let  $A$  be the number of documents in this set. Further, let  $R_a$  be the number of documents in the intersection of the sets  $R$  and  $A$ .

The recall and precision measures are defined as follows.

- Recall is the fraction of the relevant documents (the ser  $R$ ) which has been retrieved i.e.,

$$\text{Recall} = \frac{Ra}{R}$$

- Precision is the fraction of the retrieved documents (the set A) which is relevant i.e.,

$$\text{Precision} = \frac{Ra}{A}$$

Recall and precision, as defined above, assume that all the documents in the answer set A have been examined. However, the user is not usually presented with all the documents in the answer set A at once. Instead, the documents in A are first sorted according to a degree of relevance. The user then examines this ranked list starting from the top documents. In this situation, the recall and precision measures vary as the user proceeds with his examination of the answer set A. Thus, proper evaluation requires plotting a precision versus recall curve as follows.

As before, consider a reference collection and its set of example information requests. Let us focus on a given example information request for which a query q is formulated. Assume that a set Rq containing the relevant documents for q has been defined without loss of generality; assume further that the set Rq is composed of the following documents

$$Rq = \{ d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123} \}$$

Thus, according to a group of specialists, there are ten documents which are relevant to the query q.

Consider now a new retrieval algorithm which has just been designed. Assume that this algorithm returns, for the query q, a ranking of the documents in the answer set as follows.

Ranking for query q:

1.  $d_{123}$  \*
2.  $d_{84}$
3.  $d_{56}$
4.  $d_6$
5.  $d_8$
6.  $d_9$  \*
7.  $d_{511}$
8.  $d_{129}$ ,
9.  $d_{187}$
10.  $d_{25}$  \*
11.  $d_{38}$
12.  $d_{48}$
13.  $d_{250}$
14.  $d_{113}$
15.  $d_3$ \*

The documents that are relevant to the query q are marked with a bullet after the document number. If we examine this ranking, starting from the top document, we observe the following points. First, the document  $d_{123}$  which is ranked as number 1 is relevant. Further, this document corresponds to 10% of all the relevant documents in the set  $R_q$ . Thus, we say that we have a precision of 100% at 10% recall. Second, the document  $d_{56}$  which is ranked as number 3 is the next relevant document. At this point, we say that we have precision of roughly 66% at 20% recall. The precision at levels of recall higher than 50% drops to 0 because not all relevant documents have been retrieved. This precision versus recall curve is usually based on 11 standard recall levels which are 0%, 10%, 20, ..., 100%. For the recall level 0%, the precision is obtained through an interpolation procedure as detailed below.

In the above example, the precision and recall figures are for a single query. Usually, however, retrieval algorithms are evaluated by running them for several distinct queries. In this case, for each query a distinct precision versus recall curve is generated. To evaluate the retrieval performance of an algorithm over all test queries, we average the precision figures at each recall level as follows.

$$\bar{P}(r) = \frac{\sum_{i=1}^{N_q} P_i(r)}{N_q}$$

Where  $\bar{P}(r)$  is the average precision at the recall level  $r$ ,  $N_q$  is the number of queries used, and  $P_i(r)$  is the precision at recall level  $r$  for the  $i$ -th query.

Since the recall level for each query might be distinct from the 11 standard recall levels, utilisation of an interpolation procedure is often necessary. For instance, consider again the set of 15 ranked documents presented above. Assume that the set of relevant documents for the query  $q$  has changed and is now given by

$$R_q = \{ d_3, d_{56}, d_{129} \}$$

In this case, the first relevant document in the ranking for query  $q$  is  $d_{56}$  which provides a recall level of 33.3% (with precision also equal to 33.3% because, at this point, one-third of all relevant documents have already been seen. The second relevant document is  $d_{129}$  which provides a recall level of 66.6% (with precision equal to 25%). The third relevant document is  $d_3$  which provides a recall level of 100% (with precision equal to 20%).

### 15. 3. EVALUATION MEASURES at TREC Conference

At the TREC conferences, four basic types of evaluation measures are used: Summary table statistics, recall – precision averages, document level average, and average precision histograms. Briefly, these measures can be described as follows:

- **Summary table statistics** Consists of a table which summarises statistics relative to a given task. The statistics included are: the number of topics used in the task, the number of documents retrieved over all topics, the number of relevant documents which were effectively retrieved for all topics, and the number of relevant documents which could have retrieved for all topics.
- **Recall – precision average** Consists of a table or graph with average precision at 11 standard recall levels. Since the recall levels of the individual queries are seldom equal to the standard recall levels, interpolation is used to define the precision at the standard recall levels. Further, a non-interpolated average precision over seen relevant documents might be included.
- **Document level averages** In this case, average precision is computed at specified document cutoff values. For instance, the average precision might be computed when 5, 10, 20, 100 relevant documents have been seen. Further, the average R-precision value might also be provided.
- **Average precision histogram** Consists of a graph which includes a single measure for each separate topic. This measure is given, for instance, by the difference between the R-precision for a target retrieval algorithm and the average R- precision computed from the results of all participating retrieval systems.

#### **Self-Check Exercise**

1. What are the types of evaluation measures are used at TREC conferences?

**Note:**

**i). Write your answer in the space given below.**

**ii). Check your answer with the answers given at the end of this Unit.**

---

---

---

---

---

---

---

---

---

---

#### **15. 4. Retrieval effectiveness measures**

A viable information retrieval system must be effective in returning documents in response to an information need. While this generally means that most of the documents retrieved in response to the query should be judged by the user to be appropriate to the information need, such a vague statement of effectiveness provides no solid basis for determining how good a given system is, or for comparing one retrieval system to another.

##### **15. 4.1 BINARY VERSUS N-ARY MEASURES**

There are generally two steps in translation an information need into a query that a given retrieval system can handle. The first of these is to formulate a question that corresponds to the information need. Sometimes this is simple, but often asking the right question is itself a difficult task. Suppose, for example, that the user wishes to select a person for a given position from among a pool of candidates. This may be in any of several contexts: electing a public official, hiring a staff member, choosing a physician, and so forth. What question or questions should be asked to elicit the Information needed

to make an informed and intelligent decision? Often it is only by hindsight that the user discovers that the correct question has not been asked, that is, that the information need has not been properly defined.

The second step is to transform the question into a query that is suitable for a given information system. Some information systems can handle natural language questions directly, so that no transformation is needed. Others may require that the query be a logical expression of a list of weighted key terms, use a particular vocabulary, or have some other specific form. If the information system can handle the question directly, then the user must assume that the system can correctly parse the question and understand both its semantics and its pragmatics – a tall order. If the user must transform the question in some way, then she is faced with task of ensuring that the transformed question still correctly reflects the information need. Are the logical connectives correctly user? Do the weights really convey the proper sense of importance of the various terms? Is the word chosen from the restricted vocabulary an adequate and accurate representation of the word that the user would have preferred?

**Self-Check Exercise**

**Note:**

- i). Write your answer in the space given below.**
- ii). Check your answer with the answers given at the end of this Unit.**

---

---

---

---

---

---

---

---

## **15. 5. Evaluation of Retrieval Effectiveness**

### **15. 5.1 System Components**

Before a detailed examination of the evaluation parameters can be made, it is necessary to consider briefly the components of an information system and the system environment to determine how system performance is affected by the system environment and operations. The following systems components are of concern: acquisition and input policies, physical form of input, organisation of the search files, indexing languages, indexing operation, representation of the information items, question analysis, search, and form of presentation of the output.

Parameters related to the input policies include the error rates and time delays experienced in introducing new items into the collection, the time lag between receipt of a given item and its appearance in the file; and the collection coverage, that is, the proportion of potentially relevant information items actually included in the file. The physical input form, including document format and document length – title, abstract, summary, or full text – immediately affects the indexing and search tasks, as well as the system economics; and the organisation of the search files impacts the search process, the response time the effort needed by system operators, and possibly also the system effectiveness.

The indexing language consists of the set of available terms and the rules used to assign these terms to documents and search requests. During the indexing process terms appropriate for the representation of document content are chosen from the indexing language and assigned to the information items in accordance with established indexing rules. Among the parameters that take on special significance in this connection are the exhaustiveness and specificity of the indexing language. An exhaustive indexing language contains terms covering all subjects areas mentioned in the collection; correspondingly, an exhaustive indexing product implies that all subject areas are properly reflected in the index terms assigned to the documents. A specific index language never covers distinct subjects by using a single term, the terms used being narrow and precise.

Retrieval system performance is often measured by using recall and precision values, where recall measures the ability of the system to retrieve useful documents, while precision conversely measures the ability to reject useless materials. A high level of indexing exhaustively tends to ensure high recall by making it possible to retrieve most potentially relevant items; at the same time precision may suffer because some marginally relevant items are likely to be retrieved also when many different subject areas are covered by the index terms. When highly specific index terms are used, the precision is expected to be high, since most retrieved items may be expected to be relevant; the converse is true when very broad or general terms are used for indexing purposes because broad terms will not distinguish the marginal items from the truly relevant ones. Thus to obtain high recall an exhaustive indexing is useful in conjunction with an indexing language that provides a variety of approaches to cover the given subject area. To ensure high precision, a highly specific indexing language should be used, and the terms should carry additional content indications such as term weights and relation indications to other terms.

Assuming that the indexing is performed manually by trained persons, the variables affecting indexing operation relate not only to the exhaustivity of the indexing and the specificity of the assigned terms, but also to interindexer consistency, the influence of indexer experience on performance, and the accuracy of the assigned terms.

The question analysis and search operations are difficult to characterize. The assignment of terms from the indexing language to information requests, the formulation of meaningful Boolean statements, and the comparison of analysed requests with stored information are all complicated tasks. In principle, the content analysis operations are the same for documents and search requests, in the sense that the notions of exhaustivity and specificity are equally as applicable to queries as to documents. Thus, exhaustive query indexing using highly specific terms should produce maximum search recall and precision. In practice, the query processing is often quite distinct from the document indexing because the user is necessarily directly involved in the former but not the latter. In many systems, the query analysis and search operations are therefore delegated to

trained experts using appropriate input from the users. Document input, on the other hand, is invariably handled without user input.

The search operations are also hard to measure using objective parameters because the role of the users is not well defined in many query-formulating environments. Users are rarely asked to state recall or precision requirements, or to evaluate the output products. Yet search strategies need to be devised that respond to the users' specific recall and precision requirements. Among the characteristics that should be included in a measurement of search performance are the type of file organisation used, the type of query-document comparison in use, the effect of the search strategy on system response time and on search performance, and the relevance standards of the system users.

The form of presentation of the output is the physical representation of documents found by the system in response to the user's query. The appearance of then output affects the amount of user effort needed to look at the search results and the eventual satisfaction derived from a search. The more complete the form of the output, the easier is the relevance assessment task for the user. On the other hand, as the output is expanded from simple document numbers to full document texts, the time needed to examine the search results also increases.

The foregoing discussion makes clear that the components and parameters of the retrieval system affect the system operations and hence the evaluation results. Each component can be examined separately, or one can compare one entire system with another. In this case the parameters associated with each system are accepted as constant elements in the evaluation. However, one must understand that each of the parameters has an effect on the system, and the importance of each parameter cannot be asessed without taking into account the purposes for which the system is used.

## **15. 6. EVALUATION VIEWPOINTS AND THE RELEVANCE PROBLEM**

Information systems may be examined either from the viewpoint of the users or from the viewpoint of system operators and managers. The system managers may be assumed to include all those who influence the policy r the finances of the system, or

who are responsible to assume that an information system exists to meet the needs of its users, the effectiveness criteria of interest to the managers are not unlike those of the users. In particular, the system should meet the user requirements, and failures in the retrieval of relevant materials or in the rejection of nonrelevant items should be minimised. In addition, the managers and to some extent the users are also concerned with the costs and benefits of the system.

Among the many possible evaluation criteria of concern in the user population, six have been identified as critical.

1. The recall, that is, the ability of the system to present all relevant items.
2. The precision, that is, the ability to present only the relevant items.
3. The effort, intellectual or physical, required from the users in formulating the queries, conducting the search, and screening the output.
4. The time interval which elapses between receipt of a user query and the presentation of system responses
5. The form of presentation of the search output which influences the user's ability to utilise the retrieved materials
6. The collection coverage, that is the extent to which all relevant items are included in the system.

### **15. 7. Trends and Issues**

A major trend today is research in interactive user interfaces. The motivation is a general belief that effective retrieval is highly dependent on obtaining proper feedback from the user. Thus, evaluation studies of interactive e interface will tend to become more common in the near future. The main issues revolve around deciding which evaluation measures are most appropriate in this scenario. Furthermore, the proposal, the study, ad the characterisation of alternative measures to recall and precision, such as the harmonic mean and the E measures, continue to be of interest.

## 15. 8. SUMMARY

In this unit we have discussed the origins of research in information retrieval systems (IRS), system and user centered information retrieval measures. During the later half of twentieth century, some significant studies and experiments were conducted in the field of Information Retrieval. Important among them include Cranfield indexing experiments I and II; the Text REtrieval Conference (TREC) on effectiveness of information retrieval systems etc. The significance and findings of these studies have been reported in the unit. You find the discussion on 'Recall' and 'Precision' interesting and useful for measuring the efficiency of IRS. Further the unit contains a brief discussion on retrieval effectiveness, evaluation-viewpoints and current trends in information retrieval research.

## 15. 9. ANSWERS TO SELF CHECK EXERCISES

1. **At the TREC conferences, four basic types of evaluation measures are used: Summary table statistics, recall – precision averages, document level average, and average precision histograms.**

### 15.10 Glossary of Terms

**Information Retrieval:** The location and presentation to a user of information relevant to an information need expressed as a query

**Effectiveness** : The quality of the information system response to the Information need.

**Precision** : The proportion of retrieved documents that are relevant

**Recall** : The proportion of relevant documents that are retrieved.

### 15.11 RECOMMENDED BOOKS FOR READING

1. Baeza-Yates, Ricardo and Ribeiro-Neto, Berthier. Modern information retrieval. – Delhi: Pearson Education, 2004

2. Chowdhury, G. G. Introduction to modern information retrieval. – London: Library Association Publishing, 1999 (especially chapters 8, 16 and 17)
  3. Ellis, David. Progress and problems in information retrieval. London: Library Association Publishing, 1996
  4. ICKERY, B.C. 1970 Techniques of information retrieval. London: Butterworths, 1970.
  5. KEY papers in the design and evaluation of information systems/ edited by Donald King. New York: Knowledge Industry Publications, Inc., 1978.
-

---

**Unit -16****RETRIEVAL EFFICIENCY & EVALUATION STUDIES**

---

**OVERVIEW OF THE CONTENTS**

16.0 OBJECTIVES

16.1 INTRODUCTION

16.2 EFFICIENCY MEASURE

16.3 RELEVANCE

16.4 RECALL

16.5 PRECISION

16.6 RECALL AND PRECISION DEVICES

16.7 FACTORS AFFECTING RECALL AND PRECISION

16.8 RETRIEVAL PERFORMANCE EVALUATION

16.9 EVALUATION OF SYSTEM COST AND EFFICIENCY

16.10 EVALUATION STUDIES

16.11 EVALUATION CRITERIA

16.12 EVALUTION STEPS AND EVALUATION STAGES

16.13 FAIRS

16.14 MEDLARS

16.15 SMART

16.16 COMPARISON OF SMART WITH MEDLARS

16.17 CONIT

16.18 EVALUATION OF CONIT

16.19 SUMMARY

16.20 ANSWERS TO SELF CHECK EXERCISES

16.21 GLOSSARY OF TERMS

16.22 RECOMMENDED BOOKS FOR READING

## 16.0 OBJECTIVES

- To brief you about the concept of efficiency of the retrieval system
- To help you understand the basic concepts such as relevance, recall and precise and their relationship (recall and precise)
- To identify the criteria for measuring the performance of IRS and how these factors many affect the performance
- To brief the measures of system cost and efficiency

### 16.1. RETRIEVAL EFFICIENCY: INTRODUCTION

Efficiency measured evaluates how much effort or system resources must be expanded in identifying and retrieving relevant documents. The ideal IR system should only retrieve those items deemed relevant, with a minimum of overhead. Often there is a tradeoff, where effectiveness takes priority over efficiency. Early IR system evaluation efforts relied on system centred measures focused on retrieval performance, with user consideration being secondary or non-existent.

Today user-centred measures for IR system design and evaluation have become important considerations, and such measures include quantitative assessment of system usability as well as user affective response to a system.

The efficiency of the retrieval clearly depends on the index structure used to solve the nearest-neighbor search problem. Due to the linear dependency of the execution costs on the number of dimensions and the number of data items, the efficiency of features is simply given by their dimensionality. Finally, we applied our benchmark for a large number of features coming from two different retrieval systems. The features used in CHARIOT [The00] are based on color moments, texture moments and color histograms.

The features of [Hec00] apply the wavelet transformation to represent the color distribution of images. The CHARIOT system further supports the combination of features and the partitioning of images. As such, it offers a large number of feature combinations. Given the benchmark, we have determined for each basic feature type and

each feature combination an effectiveness value and an efficiency value. Drawing these values in a two-dimensional plot, we are able to easily relate different features and combination of features according to their effectiveness (which is the best combination of features) and their efficiency (which feature allows for a fast retrieval with a relatively good quality).

## 16. 2. EFFICIENCY MEASURE

The efficiency is the second important measure of our benchmark. The retrieval costs obviously depend on the index structure with which the benchmark solved the nearest-neighbor search problem. Recent work has shown that search costs in high-dimensional spaces are exponentially dependant on the dimensionality of the features [BBKK97, WSB98, BGRS99]. As such, it becomes obvious that above some dimensionality threshold all data items must be considered to answer the query [WSB98]. Rather surprisingly, this is not only a theoretical phenomenon: in as low as 10-dimensional feature spaces for images, a brute-force sequential scan often performs better than a hierarchical organization of the data set. Newer approaches like the VA-File [WSB98], the IQ-Tree [BBJ 00] or the P-Sphere Tree [GR00] perform better than the sequential scan, but are still linear dependent on the number of dimensions and the number of data items. Consequently, the (total) number of dimensions directly determines the retrieval efficiency of the feature. Absolute response times for the retrieval, however, further depend on the index structure that performed the search and the database size. Our implementation of the benchmark uses the VA-File [WSB98] to search for similar images. A nice property of the VA-File is that it can combine different features on the fly resulting in a still linear dependency on the number of dimensions. Other approaches like Fagin's A0-algorithm [Fag96] suffer from a more than linear dependency.

### Effectiveness- efficiency dependence

An effectiveness and efficiency value to each feature and feature combination. As motivated above, we use *EFF* as the effectiveness measure and the dimensionality as the efficiency measure. To compare the effectiveness-efficiency dependence of different

features, plot the values in a two-dimensional diagram. The dimensions represent the dimensionality of the feature and its effectiveness, respectively.

### 16. 3. RELEVANCE

The concept of relevance is centred to IR (Sarasevic , 1975). *In an* system, items retrieval may or may not be relevant, although they meet the criteria for retrieved./ Due to the imprecision of language, not all items retrieved will be equally relevant to the users need. Many evaluation measures ultimately rely on the concept of relevance. This is complicated by the fact that relevance itself may be measured in different ways. For instance, relevance may be assessed based on the user information need or the query submitted (Korfthage, 1997). The user may make document relevance judgments, or these may ultimately be made by the system evaluators for given information tasks. Document is often judged as relevant or not there are varying degrees of relevance. The degree of relevance may change if one considers the possibility of changing information needs as the users peruse documents. Therefore relevance may be situational under some circumstances (Wilson, 1973). Concepts related to relevance may also be applied to the evaluation. *Pertinence measures* how well a document matches an information need. *Usefulness* is another factor which relates to how well a document may serve a different information need, or the novelty that information within a document provides. Hence, relevant documents whose contents are already known may not be considered useful. These are user centred measures.

Whenever one talks of information retrieval, one term, which assumes significance, is “relevance”. But then, the question “What is relevance” is difficult to answer. This is because, for relevance there are degrees. In other words, ‘yes – no’, ‘either – or’ approach to relevance in index term assignment is difficult. Thus, a term can be assigned to a document on a weighted basis, the amount of weight depending on the probability that a searcher using that term would find that document relevant. Maron called this new approach as “probabilistic indexing”.

The idea of measuring relevance spread to the Cranfield project (which would be discussed in the latter sections). However, the concept of relevance as viewed at Cranfield was different from that of Maron. On the other hand, Golfman was suggesting that relevance was not intrinsically a measure, because it is not additive. So, the question turned on not to “why relevant”, rather “how relevant”. But, in 1967, Cuadra and Katter published a paper entitled “Opening the Black Box of Relevance” wherein they reported the effects of varying the questions given to relevance judges by informing them what the retrieved documents are used for.

For example

Use in stimulating ideas, creative approaches etc.,

Use in relation to specific task or,

Use in preparing an exhaustive bibliography.

And the authors concluded that “...Thus, it is possible to obtain higher or lower relevance scores simply by telling judges how documents are to be used”. This is essential, as otherwise judges would make their own assumption leading to variations within the “black box of relevance”. All this led many to a question “the relevance of relevance”. Nevertheless, the concept of relevance continues to be used as can be seen from the fact that relevance judgments are sought for from users in the context of SDI (Selective Dissemination of Information).

#### 16. 4. RECALL

On the other hand, Recall is the ratio between the number of relevant documents retrieved and the total relevant in the file or system. For example, using the common frame of reference, recall ratio may be shown, as follows:

$$\text{Recall Ratio} = \frac{\text{Relevant retrieved}}{\text{Total retrieved}} \times 100 = \frac{a}{a + b} \times 100$$

For example, if there are 100 documents, relevant to a particular question in a file or system and if 60 are retrieved and 40 are missed, then the recall ratio is 60: 100, i.e., 60%. While this may be a useful measure, there is the problem of determination of the total number of relevant documents in the file / system, unless one scans the whole file / system completely. For large collections, this is ‘exceedingly impractical’. But, Salton (7) suggests a number of alternative methods for this purpose. They are:

- a) Use of sampling techniques;
- b) Designation of source documents;
- c) Scanning of condensed formats indexed by one of the “added” words in search request; and
- d) conducting a number of searches using the same search request but different search methods – word system, thesaurus, statistical, phrase, etc. further, as mentioned, there is the problem of determining relevance in the context of recall ratio.

## 16. 5. PRECISION

Precision is the ratio between the relevant retrieved and the total retrieved documents. As a matter of fact, it is the measure of the work still to be done in screening out irrelevant material at next level of search. For example, if in retrieving 8 wanted documents, one retrieves a total of 100 documents (i.e., 8 wanted and 92 unwanted), the precision ratio is 8/100 or 80%.

Similarly, if an index retrieves a total of 50 reference in response to a question, out of which 40 are relevant then the precision ratio is

$$\frac{40}{50} \times 100 = 80\%$$

In the context of evaluation, the following common frame of reference would be useful :

	Relevant	Not relevant	Total
Retrieved	A HITS	B NOISE	A + b Total retrieved
Not retrieved	C Misses	D Correctly rejected	C + d Total not retrieved
Total	A + c Relevant	B + d Not relevant	A + b + c + d Total collection

$$\text{Precision Ratio} = \frac{\text{Relevant retrieved}}{\text{Total retrieved}} \times 100 = \frac{a}{a + b} \times 100$$

It may be asked as to why irrelevant materials are retrieved in relation to a query. They are usually retrieved because they contain combination of words, terms, or phrases that matched those in the search statement. In other words, retrieval of irrelevant material is unavoidable and an accident. Further, studies have shown that parameters, such as library size and topical specificity, influence the number of irrelevant retrievals. Thus, precision cannot be used as a measure of retrieval effectiveness in libraries of different size and / or specificity.

## 16.6 . RECALL AND PRECISION DEVICES

Several devices are used to manipulate indexing to obtain optimum recall and precision. These are:

### 16.6.1 Recall devices

- Synonym control
- Word form control, stems
- Classification, including
- Hierarchies
- Lattices
- Facet analysis

Semantic factoring  
Clumps and clusters.

### **16.6.2 Precision devices**

Coordination  
Links  
Roles  
Weighting  
Relational indexing

Coverage is a measure of the completeness of a collection. It is related to recall in that it is likely to be of concern only to the user who needs a high recall. i.e. when a comprehensive search is warranted.

Novelty refers to the newness of information supplied by a service and us of oblivious importance in the evaluation of current awareness services, since, presumably, a good current awareness service will bring documents to the attention of users before they learn of them by other means.

### **Self-Check Exercise**

#### **1. What are the Recall devices?**

**Note:**

- i). Write your answer in the space given below.**
- ii). Check your answer with the answers given at the end of this Unit.**

---

---

---

---

---

---

---

---

---

---

## 16.7. FACTORS AFFECTING RECALL AND PRECISION

Thus, while Recall and Precision refers to the ability to retrieve only relevant items (or, put differently, the ability not to retrieve irrelevant items), these two performance criteria have inverse-relationship and are influenced by the following factors:

- I. Requests that imperfectly represent information needs;
- II. Indexing factors;
- III. Search strategy factors; and
- IV. Vocabulary factors.

On the basis of an analysis of retrieval errors executed in the operation of MEDLARS, Lancaster (6) described the major causes of error as follows:

---

Factors	Recall failures	Precision failures
i) Index language	Lack of specific terms (entry vocabulary) Inadequate hierarchical Cross-reference structure Roles, or other Relational indicators Causing over-preciseness	Lack of specific terms (Index terms) Defects in hierarchy False coordinations Incorrect term relationships
ii) Indexing	Lack of specificity	Exhaustive indexing

	Lack of exhaustivity Omission of important Concepts Use of inappropriate Terms	Use of inappropriate terms
iii) Searching	Failure to cover all reasonable approaches to retrieval Formulation too Exhaustive  Formulation too specific	Formulation not sufficiently exhaustive Formulation not sufficiently Specific Use of inappropriate Terms or term Combination
iv) User / System	Requests more specific than actual information needs	Requests more general than actual information needs
v) Other	Computer processing Clerical	Computer processing Clerical Value judgment Inevitable retrievals

---

**Self-Check Exercise**

**2. What are the factors affecting Recall and Precision?**

**Note:**

**i). Write your answer in the space given below.**

**ii). Check your answer with the answers given at the end of this Unit.**



experiments. Despite this tendency, laboratory experimentation is still dominant. Two main reasons are the repeatability and the scalability provided by the closed setting of a laboratory.

### **16.9. EVALUATION OF SYSTEM COST AND EFFICIENCY**

The art of efficiency analysis is not as far advanced as the analysis of system effectiveness. This is because accurate cost data in terms of time, effort and money spent are difficult to obtain, and because the value of improved information services and the benefits derivable from them is impossible to ascertain in most environments. Furthermore, when identifying information system costs, invariably one is forced to look at noncomparable situations. The cost differences between two systems, such as an automated and a manual one, may not accurately reflect the value of either system. The automated system might, for example, be used for purposes other than information storage and retrieval, or it might be usable on a 24 – hour per day basis, whereas a manual system might not. Thus an efficiency evaluation involves a great many intangible factors, which may hamper a concrete analysis and render the results unreliable or meaningless.

### **16. 10. EVALUATION STUDIES**

Evaluation is essentially a diagnostic procedure. Evaluation of indexing systems is undertaken to examine the performance of indexing systems, which are adopted in the information retrieval activities. Moreover, evaluation tends to be expensive, and can only be justified if the evaluation technique or programme is likely to lead a significant improvement in the performance of the system.

Evaluation may be defined as:

- The process of determining to what extent the objectives as actually being achieved:
- Providing information for decision-making:

- The comparison of performance with some standards to determine whether discrepancies existed; and
- The systematic investigation of the worth or merit of some objects.

## Performance

Performance is an indication of how well a service or activity performs and can be measured in terms of the input costs and output quantities, quality, timeliness, availability, accessibility, etc.

## Measure

Measure is generally used to mean any process for describing in quantitative values: things, people, events, etc. It also means the value being measured.

## 16.11. EVALUATION CRITERIA

Perry and Kent are credited for bringing the concept of evaluation into information retrieval systems in the mid-fifties. The evaluative measures, they suggested, were :

$$L / N = \text{Resolution factor} \qquad (N - L / N) = \text{Elimination factor}$$

$$R / L = \text{Pertinency factor} \qquad (L - R) / L = \text{Noise factor}$$

$$R / C = \text{Recall factor} \qquad (C - R) / C = \text{Omission factor}$$

Where, N = Total number of documents

L = Number of retrieved documents

C = Number of relevant documents

R = Number of documents that are both retrieved and relevant

However, only two measures – namely, precision (new name for pertinency), and recall are presently used in evaluation studies. In some cases, evaluators have preferred the measure fallout  $(L - R) / N - C$  to precision. In addition, one other factor, which is important in the context of evaluation, is relevance.

### 16.12. EVALUTION STEPS AND EVALUATION STAGES

There are a number of distinct steps involved in the conduct of an evaluation programme. An information service can also be evaluated at various stages in its development. The major steps involved are:

1. Defining the scope of the evaluation
2. Designing the evaluation programme
3. Execution of the evaluation
4. Analysis and interpretation of the results
5. Modifying the system or service on the basis of the evaluation results.

#### Self-Check Exercise

##### 1. What are the major steps involved in the Evaluation Stages?

**Note:**

- i). Write your answer in the space given below.
- ii). Check your answer with the answers given at the end of this Unit.

---

---

---

---

---

---

---

---

---

---

### 16.13. FAIRS

D.E Berninger and his associates evaluated the operational systems FAIRS, the retrieval system of the U.S.Federal Aviation Agency. The system contains ten thousand technical reports indexed with the aid of a thesaurus of descriptors. Whenever a specific term was used, the corresponding generic term was added to the index i.e. the method of up posting was applied. They system was put to evaluation by searching it to answer ten search request, actually selected from genuine requests previously submitted buy the users. User feedback on the relevance of the retrieved documents was collected and the precision ratios were established.

Then, two methods were used to calculate the recall ratio. In the first method, users were presented with a 10% random sample of the collection and were asked to identify the relevant documents in the sample against each test search request. This figure of relevant documents  $X_1$  was used to determine the possible number of relevant documents in the whole collection 'X' as ' $X = 10 \times X_1$ '

In the second method a supplementary test was conducted on 20 sources documents. A set of completely synthetic searches were carried out on the 20 sources documents. A search was considered successful it is recalled the source document and the recall ratio was derived on the basis of percentage of successful searches.

If a query includes generic term G1 and specific terms S1 and S2 belong to other generic terms G2, and G3 respectively.

G1 with NT S3, S4, S5 etc

G2 with NT S1, S6, S7 etc

G3 with NT S2, S8, S9 etc.

The above example shows that the two specific terms S1 and S2 belong to other generic terms G2 and G3 respectively.

Four strategies were used to derive the performance ‘figures: precision ratio, recall ratio (by first method), and recall ratio (by second method):

- Search strategy A included originally chosen search terms (whether specific or generic) i.e. G1, S1, S2.
- Search strategy B eliminated all specific terms and was limited to generic terms G1, G2 and G3
- Search strategy C included generic terms G1, G2, and G3 and eliminated all non-pertinent specific terms S3, S4, S5, S6, S7, S8, S9
- Search strategy D co-ordinated generic terms with selected specific terms G1 with any terms specific to G2 and G3, then G2 with any terms specific to G1 to G3, and so on

The results of the study are presented in the following table

Strategy	A	B	C	D
Precision ratio	59	35	38	45
Recall ratio (first method)	22	73	66	50
Recall ratio (second method)	70	90	80	75

The result indicated that efforts to raise recall resulted in the precision falls. The recall ratios derived by the two methods were not the same, yet they varied in the same way with precision.

#### 16.14. MEDLARS

A large system MEDLARS containing 70,000 biomedical article was evaluated during 1966-67. On average 6.7 subject terms were used to represent the concepts in each article. The terms were selected from a thesaurus, MeSH (Medical Subject Headings), which consists of about 7,000 main subject headings that can be supplemented by sub-headings, Hierarchical searches are supported by the system.

A selection was made from the existing user groups that could be (a) supply a certain volume of test questions; (b) cover all the kinds of requests made (categorized as on diseases, on drugs, etc.) (c) include all kinds of users (academic, research, pharmaceutical, clinical, government etc.; and (d) vary according to the degree of user/system interaction (personal interaction, positive or negative or no local interaction).

The 21 user groups so selected provided 302 fully analyzable test searches. Search output, along with photocopies of the articles were provided to the requester for evaluation using the scale: H1- of major value, H2-of minor value, W1-of no value and W2-or unknown value. Precision ratios were calculated for over all precision (H1+H2) and 'major value' precision 'H1' only.

Sampling techniques were used to establish overall recall ratio and 'major value' recall ratio for each of the 302 searches and these are then averaged to arrive at the following figures.

	<b>OVERALL</b>	<b>MAJOR VLAUE</b>
RECALL VALUE	57.7%	65.2%
PRECISION RATIO	50.4%	25.7%

Each search was analyzed in detail, and failures in recall and precision, were ascribed to the indexing language, to indexing, to user-system interaction, searching, to computer processing. This analysis leads to a series of recommendations on upgrading system performance.

In about 23% of the 302 searches, a recall failure and in about 37 of the searches a precision failure were attributed to “inadequate user-systems interaction” that resulted analyst/searchers inadequate interpretation of the information requirements of the users.

The four levels of interactions recognized are:

1. Personal interaction – the user visited a MEDLARS center and discussed his information needs personally with a system operator.
2. Positive local interaction – a local librarian discussed the information needs before transmitting the request to MEDLARS centre.
3. Negative local interaction – a local librarian simply transmitted the request.
4. No local interaction – the request mailed his request directly to MEDLARS center.

It was hypothesized that the first group of requests would given the highest performance but the results showed the interactive group 1 and 2 performed worse than the neutral groups 3 and 4.

The success of the neutral groups is due to the fact that the requester submitted his information needs in verbal form, in his own natural language, without being influenced by the logical and linguistic constraints of the MEDLARS systems, as evidenced by no interaction with the system.

The failure of the interactive groups is due to the fact that the user has initially a less well-formed idea of what he is seeking (i.e. of the scope and constraints of the search) and when this somewhat imprecise need is discussed with a search analyst, in terms MeSH, it tends to become forced into the language and logic of the system. The

final 'request' rather than representing what the user wants, represents what he thinks of the system can give him. It appears that little knowledge of the system on user part will lead of failures. Either the user should give full freedom to the searcher to analyze the search request and formulate a strategy or he should himself learn the technique of searching a system.

Lancaster commented that "It appears crucial to the success of a MEDLARS search that the requester be required to write down, in his own natural language, exactly what type of literature he is looking for.

### **16.15. SMART**

The SMART (System for Mechanical Analysis and Retrieval of Text) project was initiated in 1961 with an emphasis on fully automated procedures for the analysis, search, and retrieval of natural language texts and has become operational 1964. From its inception, the system was designed both as a retrieval tool and as a vehicle for evaluating the effectiveness of a large variety of automatic search analysis techniques.

#### **16.15.1 Design of SMART System**

SMART system consisted of three parts: an automatic content description (indexing) system; a supervisory or monitory system; and an evaluation system. the SMART indexing system was based on seven language analysis tools.

- Methods for automatically extracting important words from natural language texts of incoming user queries and documents excerpts (titles, abstracts or full texts.)
- Sophisticated suffix cut-off procedures which would be used to transform the words extracted to word stem form;
- Synonym dictionaries of thesaurus;
- Hierarchical term arrangement systems;
- Syntactic analysis systems;
- Semantic analysis systems;
- And statistical frequency analysis systems;

In SMART, a document or query is represented by a vector or terms i.e. words that carry the concept of document or query.

#### Steps in Automatic Indexing of Document or Query

1. the document text or query is broken into words;
2. High-frequency function word such as “and, of, or , but, when, where etc., are removed with the help of a stop word list.
3. Reducing each word to a word stem using suffix removal procedures broadens the scope of the remaining word occurrences.

E.g. economist, economists, economical, economically, economize, economizes, economized, economizing, economies etc. into ECONOM

4. Multiple occurrences of a given word stem are combined into a single term for incorporation into a document or query vector (frequency analysis is carried out)

E.g. document 1 dealing with fruits is represented in the following vector;

(apple, 4; pear, 3; guava, 2; plum, 1)

5. Transforming the word stem vector into useful term vectors by two manipulations; first a term weight can be assigned to each term reflecting the usefulness of the term in the collection environment; and second, terms whose usefulness is inadequate as reflected by the low term weights can be transformed into better terms.

Terms weights are assigned in the following manner:

- a) Calculating the frequency of the term in a document or query
- b) Calculating the discrimination value of the term in distinguishing the specific document from other documents in the collection.
- c) Calculating the document frequency i.e. the number of document to which the term is assigned.

$$\text{WEIGHT} = \frac{\text{TERM FREQUENCY}}{\text{DOCUMENT FREQUENCY}}$$

## DOCUMENT FREQUENCY

### OR TERM FREQUENCY x DISCRIMINATION VALUE

6. Terms whose weight is neither too large nor too small are incorporated directly into the document or query vector.

Terms whose weight exceed a given threshold level are considered too broad and unspecific. These are rendered more specific by being combined with other terms into term phrases that contain two word stems before including into the vectors.

e.g. people in need of information require effective retrieval services (original sentence)

PEOPLE INFORM EFFECT RETRIEV SERVICE      word stems  
PEOPLE INFORM    EFFECT RETRIEV    INFORM EFFECT    term  
EFFECT SERVICE    INFORM RETRIEV    RETRIEV SERVICE    phrases  
INFORM SERVICE

In forming term phrases, the indexing system follows certain rules such as the word distance should not exceed 4 and theses should be in a single sentence etc.,

Terms with low weight are considered too specific and are broadened by grouping them into term classes similar to a thesaurus entry. The thesaurus class identifiers are then incorporated into document and query vectors, instead of the individual rare terms.

Document in SMART system are automatically classified based on vectors and placed in clusters where items that appear reasonably similar to each other are placed in close proximity.

The supervisor or monitoring system could process the query and document vectors calling necessary language analysis tools and by suitable matching operating, could supply to the user, references to those documents whose content vector appeared to be similar to the corresponding query vectors. The evaluation system provides formal assessments of system effectiveness in terms of satisfaction of users.

## **16. 15.2 Search Process in SMART**

Information is retrieved by a complete vector matching method providing for each query-document pair a coefficient of similarity. A ranking is obtained for the stored items (i.e. 100% similar, 90% similar etc.,) in decreasing order of similarity, and available number of documents is presented to the user. This permits the user to consider first those documents which appeared to the system to be most similar to the specified query and select some relevant documents.

A new search operation could then be initiated by automatically altering the initial query vector so as to retrieve more documents similar to the documents considered relevant by the user. A number of such relevance feedback operations could be carried out so that users information needs could be met satisfactorily.

## **16. 15.3 Evaluation of SMART**

A number of tests conducted in varied subject areas like engineering aerodynamics and documentation indicated that:

- the order of merit is generally the same for all three subject areas
- the use of un weighted terms is less effective than the use of weighted terms
- the use of document titles alone is always less effective for content analysis purpose than the use of abstract
- the thesaurus processes involving synonym recognition perform more effectively than the word stem extraction method, where synonyms and other word relations are not recognized.
- The thesaurus and statistical phrase methods are substantially equivalent in overall system performance

- Other dictionaries including term hierarchies and syntactic phrases exhibited poor performance.

#### **16.16. COMPARISON OF SMART WITH MEDLARS**

One of the aims of the SMART project had been the comparison of fully automated text processing systems with the manual indexing systems like MEDLARS. MEDLARS system is based on a manual analysis of documents and incoming search requests.

A sub-collection from the full MEDLARS collection and a subset of original queries submitted to MEDLARS were used for a comparative study, by processing both as per SMART methodology. The recall and precision results, averaged over 29 queries, exhibited recall ratios about 40% lower for SMART, than for MEDLARS; The precision loss was between 30 to 40%, where SMART used its standard word stem extraction method only. When the ranked outputs provided by SMART were used, the situation improved drastically, i.e. a deficiency of only 16% in recall and 19% in precision. Through the use of relevance feedback methods, this 15-20% deficiency in recall and precision turned into an advantage of 4-7% after one feedback operation, and of 10-13% for two feedback iterations, over MEDLARS.

To improve the effectiveness of the SMART system an automatically generated dictionary that contains a list of all terms in decreasing order of term discrimination value, designed to exclude all high frequency terms was used. The automated dictionary provided a 10% improvement in recall and a 20% in precision, over the standard word stem extraction methods. With the use of SMART thesaurus, improvements of about 25% in average recall and precision ratios were achieved over the standard word stem process.

The SMART-MEDLARS comparison can be concluded as follows:

- 1) The strong points of the automatic retrieval system appear to be the vector matching techniques which furnish ranked document output, the automatic construction methods for word control lists, and the feedback operations.

- 2) The simple word stem extraction process using document abstracts and query texts is only 15-20% less effective than the best available manual indexing based on controlled vocabularies.
- 3) Automatic language normalization procedures can be used to build dictionaries and thesauri, whose operations produce output results equivalent to standard manual indexing.

Several heterogeneous retrieval systems are commercially available and there are significant differences among these systems in terms of command languages, search aids, (thesaurus etc.) provided by them. The end-user has to learn the syntax and semantics of each of the information retrieval system if he wishes to exploit its capabilities. Expert systems based interfaces like CANSEARCH are designed and implemented but they are confined to specific subject areas and available on specific systems only.

### **Self-Check Exercise**

#### **1. How many parts in SMART system?**

**Note:**

- i). Write your answer in the space given below.**
- ii). Check your answer with the answers given at the end of this Unit.**

---

---

---

---

---

---

---

---

---

---

## **16.17. CONIT**

CONIT is a general-purpose interface aimed at connecting three different commercially available retrieval systems (MEDLINE, SEC ORBIT AND DIALOG), which together contain 300 databases by 1983. CONIT connects the user to these three systems but presents to the user what appears to a single, common (virtual) system by allowing user requests in a common command language. These requests are in turn translated by the interface into appropriate commands acceptable to the host system selected by the user. The interface provides instructions and additional search aids to help the novice user.

### **16.17.1 Automated Keyword / Stem Searching**

The problem of effective searching by a novice user across database with heterogeneous indexes was met by a natural language; free vocabulary approach to searching that emphasizes the use of keyword stems as the basis for searching.

A search on the topic 'transplantation rejection' results in two words stems 'transplant and reject'. Then CONIT conducts truncated searches on each of the stemmed forms in all the indexes that can be searched with a single command in the connected database. The sets retrieved for each individual sub search are then combined with a Boolean OR and finally these separate unions are combined with the Boolean AND operator to yield the resultant set.

Searching in an information retrieval system is not limited to subject searching only. Users may wish to search the system on personal names, but the rendering of personal names varies from system to system. e.g. Lancaster, F.W. in one system and Lancaster, F.W in another system.

To get over this problem, the user is permitted to request personal names searches in a common format. CONIT then translates this format into the one appropriate for

the database being searched – correct spacing and punctuation between entry element and CONIT supplied the initials.

CONIT names the sub searches and the resultant search and reports to the user the number of documents in each set. All this is done automatically without user intervention. If any of the sub searches yield null results. CONIT suggests browsing the index terms or the thesaurus around the non-responsive term. If a truncated search causes a search buffer overflow. CONIT replaces truncated search with an exact match, full-word search or full-phrase search.

### **16.17.2 Search History and Reconstruction**

CONIT system has a search history recording and reconstruction capacity. For each search CONIT records the full search formulation, the database system used, the number of documents found in the resultant set or in any component sets formed in creating the resultant set, and the set names as given by CONIT and by the retrieval system, and whether the set is currently available in the retrieval system etc. all this information will be available on-line to the end-user.

When a user requests any component or compound search formulation to be repeated in any other database or set of databases, CONIT refers to the search history and repeats the search, after connecting to appropriate systems and databases. The dropped or no longer available sets, in the original request are generated first before any operation (output generation or combining sets) is performed, totally transparent to the end-users.

### **16.18 EVALUATION OF CONIT**

Some 16 end-users selected from different levels (two medical doctors, one non-academic university staff, two professors, one post-doctoral fellow, 6 graduate students, and 4 under graduates), none of whom previously operated either CONIT or any one of the connected retrieval systems, performed searched on 20 different topics using CONIT

with no assistance other than that provided by the interface. These same users performed searches on the same topics with the help of a human expert who searched the retrieval systems directly.

The parameters considered for the experiments include: total search time, the time spent by the users in getting help from CONIT, the actual search time which includes the time spent in issuing commands and getting their responses, the time spent in displaying retrieved records and assessing their relevance, the number of relevant documents retrieved, the estimated number of relevant documents in the documents, and the number of databases searched.

The results indicated that:

- Sometimes CONIT and sometimes the human expert were clearly superior in terms of search effectiveness i.e. recall and precision ratios. In general, however, end-users searching alone with CONIT achieved higher on-line recall at the expense of longer search sessions.
- In terms of search time which has a direct bearing on the cost of search, particularly in the case of commercial databases, experts had spent at least 20% less time than their counterparts, end-users.
- The number of relevant records found and viewed on-line was much higher for the user CONIT sessions than for the human expert sessions; this led to longer search sessions in case of user CONIT sessions.
- Human experts seemed to be more sophisticated, complex and comprehensive in their search i.e. they used all possible tools to raise recall and precision.
- Human experts regularly took advantage of such precision-enhancing devices as proximity searching, important term searching and subheadings and other controlled vocabulary searching.
- Experts also used such recall-enhancing devices as truncated searching and searching on all more specific terms given for a term etc.
- End-users did not make use of these devices, as they are not aware of their availability in the system.

- A number of end-users used the facility of browsing indexes and thesaurus using terms found in document records, as CONIT suggests this during execution.

It is concluded that advanced experimental intermediary techniques are capable of providing search assistance whose effectiveness is at least similar to that of human experts in some contexts.

### **16.19. SUMMARY**

In unit -17 we have discussed the information retrieval system effectiveness. This unit discusses the efficiency measures of information retrieval system- (IR's) with emphasis on recall and precision ratio. You will also find a useful discussion synonym control and classification system, which contribute significantly in the enhancement of retrieval efficiency. The concept of co-ordination and relational indexing in this unit makes your understanding of the subject a more exhaustive. Other factors affecting the recall and precision ratio of information retrieval systems such as index language, indexing, searching, user etc have been dealt in the Unit appropriately. The unit concludes by discussing the cost and efficiency measures.

Let us recapitulate what has been discussed so far in this unit

- A number of experiments and case studies have been carried out to evaluate the effectiveness of information storage and retrieval systems.
- The first study, popularly known as Cranefield Project I, compared the efficiency of four indexing systems, while in the second phase (Cranefield Project 2 ) a number of variables studied.
- Evaluation of operational systems (FAIRS and MEDLARS), and experimental systems (SMART) have been discussed.
- A user interface has been devised, as an intermediary system, to help the end-users without any experience on searching by various information retrieval systems. The evaluation of interfaces, CANSERCH and CONIT has been discussed.

### **16.20 ANSWERS TO SELF CHECK EXERCISES**

1. The Recall devices are Synonym control, Word form control, stems; Classification,

including, Hierarchies, Lattices, Facet analysis, Semantic factoring, Clumps and clusters.

2. The factors affecting the Recall and Precision are Requests that imperfectly represent information needs; Indexing factors; Search strategy factors; and Vocabulary factors;.
3. The major steps involved are: Defining the scope of the evaluation; Designing the evaluation programme; Execution of the evaluation; Analysis and interpretation of the results and Modifying the system or service on the basis of the evaluation results.
4. SMART system consisted of three parts, the parts are: an automatic content description (indexing) system; a supervisory or monitory system; and an evaluation system

## **16.21 GLOSSARY OF TERMS**

### **Information Retrieval**

The location and presentation to a user of information relevant to an information need expressed as a query.

### **Precision**

The proportion of retrieved documents that is relevant

### **Recall**

The proportion of relevant documents that are retrieved.

### **Search History**

A mechanism for tracking the history of a user session or of a collection of user sessions. The search history should show what the available choices are at any given point, what moves have been made in the past, short-term tactics, and annotations on choices made along the way.

### **Efficiency**

S

A measure of parallel algorithm performance given by  $\phi = \frac{S}{N}$ , where S is Speedup and N is the number of processors.

### **Precision**

Set of syntactic features that describe the text segments to be matched, ranging from simple words to regular expressions.

### **Recall**

An Information retrieval performance measure that quantifies the fraction of known relevant documents, which were effectively retrieved.

## **16.22 RECOMMENDED BOOKS FOR READING**

1. Ellis, David. Progress and problems in information retrieval. London: Library Association Publishing, 1996
2. Baeza-Yates, Ricardo and Ribeiro-Neto, Berthier. Modern information retrieval. – Delhi: Pearson Education, 2004
3. Chowdhury, G. G. Introduction to modern information retrieval. – London: Library Association Publishing, 1999
4. VICKERY, B.C. 1970 Techniques of information retrieval. London: Butterworths, 1970.
5. KEY papers in the design and evaluation of information systems/ edited by Donald King. New York: Knowledge Industry Publications, Inc., 1978.
6. BORGMAN, C.L., Case, D.O. and Meadow, C.T. “The design and evaluation of a front-end user interface for energy researchers”. Journal of the American Society for Information Science 40(2): 1989. p.99-109.
7. CHARAMELLA, Y and B.Defude. A prototype of an intelligent system for information retrieval: IOTA. Information Processing & Management 23(4): 1987. p.285-303.
8. ELLIS, D. New horizons in information retrieval. London: Library Association, 1990.

9. HANCOCK-BEAULIEU, M . Fieldhouse, M and Do,T. “An evaluation of interactive query expansion in an online library cataloger with a graphical user interfaces”*Journal of Documentatilon* 51(3): 1995. p.225-243.
10. KEY papers in the design and evaluation of information systems/ edited by Donald King. New York: Knowledge Industry Publications, Inc., 1978.
11. MARCUS, R.S. “An experimental comparison of the effectiveness of computers and humans as search intermediaries”. *Journal of the American Society for Information Science* 34(6): 1983. p.381-40.
12. MEADOW, C.T., Wang, J and Yuan, W. “A study of user performance and attitudes with information retrieval interfaces”. *Journal of the American Society for Information Sciences* 46(7): 1995. p.490-505.
13. POLLIT, S. 1987. “CANSEARCH: an expert systems approach to document retrieval”. *Information Processing & Management* 23(2): 1987. pp. 119-138
14. SALTON, G and McGill, M.J. *Introduction to modern information retrieval* Aucklan: McGraw-Hill International Book Co., 1983.
15. VICKERY, B.C and Vickery, A. *Information science in theory and practice*. London: Butterworths, 1987.
16. Vickery, B.C. 1970 *Techniques of information retrieval*. London: Butterworths, 1970.