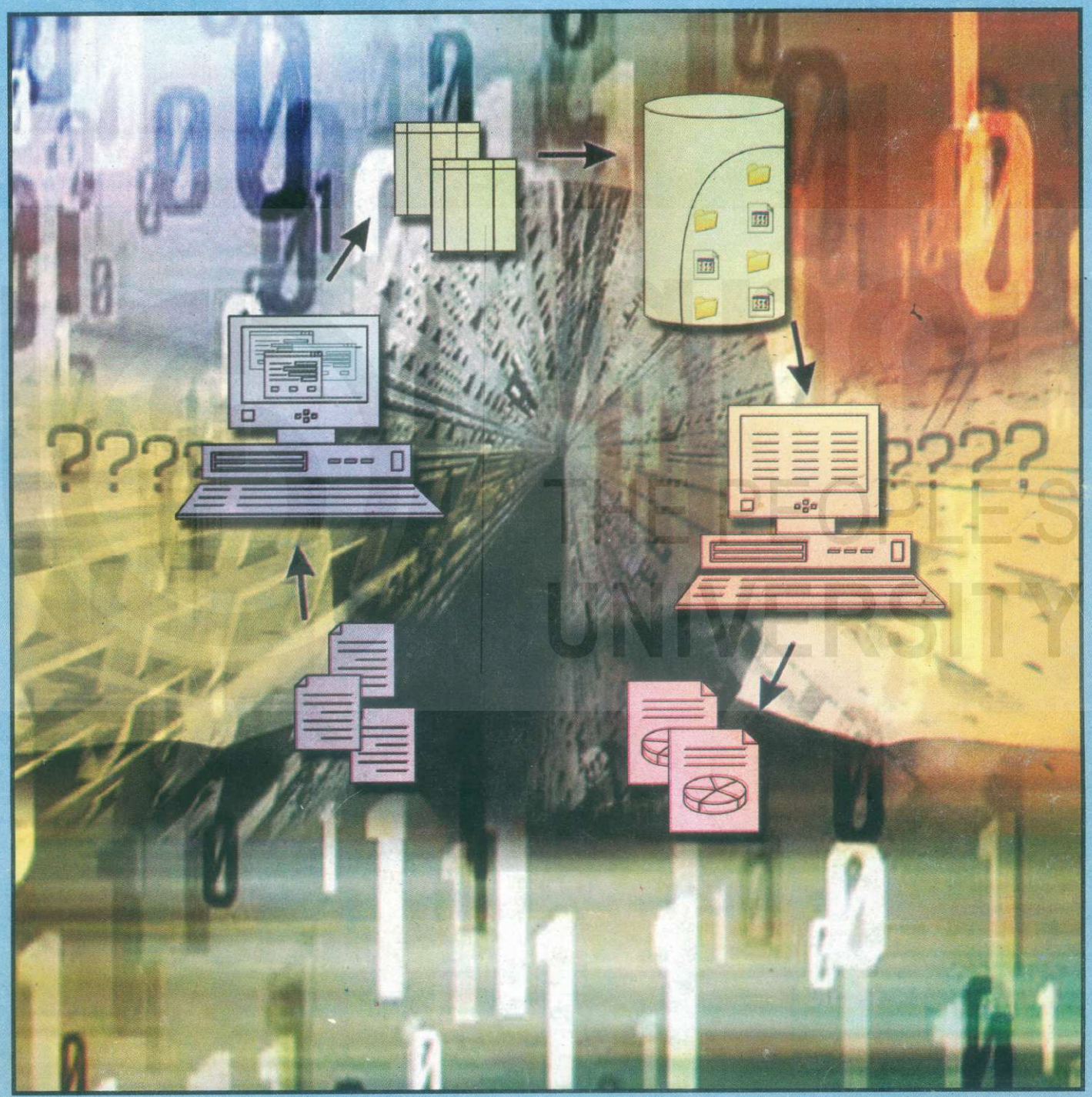
Indira Gandhi National Open University
School of Social Sciences

MLII - 102

Master's Degree in Library and Information Science (MLIS)



INFORMATION PROCESSING AND RETRIEVAL

PART I

"शिक्षा मानव को बन्धनों से मुक्त करती है और आज के युग में तो यह लोकतंत्र की भावना का आधार भी है। जन्म तथा अन्य कारणों से उत्पन्न जाति एवं वर्गगत विषमताओं को दूर करते हुए मनुष्य को इन सबसे ऊपर उठाती है।"

— इन्दिरा गांधी

ignou
THE PEOPLE'S
UNIVERSITY

"Education is a liberating force, and in our age it is also a democratising force, cutting across the barriers of caste and class, smoothing out inequalities imposed by birth and other circumstances."

– Indira Gandhi

Indira Gandhi
National Open University
School of Social Sciences

MLII-102 Information Processing and Retrieval

Introduction 5

Part I

Block 1 Organisation of Information

Unit 1: Intellectual Organisation of Information 9

Unit 2: Indexing Languages – Part I: Concepts and Types, Subject Headings Lists and Thesauri 33

Unit 3: Indexing Languages – Part II : Classification Schemes 56

Unit 4: Indexing Systems and Techniques 89

Unit 5: Evaluation of Indexing Systems 177

Block 2 Bibliographic Description

Unit 6: Principles and Evolution of Bibliographic Description 211

Unit 7: Rules for Bibliographic Description 247

Unit 8: Standards for Bibliographic Record Format 277

Unit 9: Metadata : MARC21-856 Field, Dublin Core, TEI 296

Part II

Block 3 Content Development

Unit 10: Norms and Guidelines for Content Development 331

Unit 11: Introduction to HTML and XML 352

Unit 12: Web Based Content Development 378

Unit 13: Multilingual Content Development (Using Unicode) 394

Block 4 Information Storage and Retrieval Systems

Unit 14: ISAR Systems : Objectives, Types, Operations and Design 413

Unit 15: Compatibility of ISAR Systems 433

Unit 16: Intelligent Information Retrieval Systems 448

Block 5 Information Retrieval

Unit 17: Information Retrieval Processes and Techniques 465

Unit 18: Information Retrieval Models and Their Applications 497

Unit 19: Search Strategies, Processes and Techniques 516

ignou THE PEOPLE'S UNIVERSITY

Programme Design Committee (Original) (1992)

Prof. Pandav Nayak (*Chairman*)
Prof. (Ms.) A.K. Anand
Prof. J.C. Binwal
Prof. M.A. Gopinath
Prof. B. Guha
Prof. S.R. Gunjal
Prof. S.G. Mahajan
Mr. N.M. Malwad
Dr. S.S. Murthy
Prof. K.S. Raghavan
Prof. T.N. Rajan
Prof. A.P. Srivastava
Prof. T. Viswanathan
Dr. (Ms.) Neela Jagannathan *(Special Invitee)*
Dr. R. Satyanarayana (*Convener*)

Internal Members

Dr. Uma Kanjilal
Ms. Neena Talwar Kanungo

Course Preparation Team (1992)

Prof. M.A. Gopinath
Prof. A.A.N. Raju
Dr. S. S. Murthy
Dr. M.P. Satija

Course Editor: Dr. R. Satyanarayana

Programme (Curriculum) Revision Committee (2003)

Prof. S.B. Ghosh (*Chairman*)
Faculty of Library & Information Science
School of Social Sciences
IGNOU, New Delhi

Dr. Jagdish Arora
Librarian
Indian Institute of Technology
Mumbai

Prof. S.R. Ganpule
Dept. of Library & Information Science
Yaswant Rao Chavan Maharashra Open University, Nasik

Prof. B. Guha
Professor (Retd.)
Dept. of Lib. & Inf. Sc.
Banaras Hindu University
Banaras, U.P.

Prof. Uma Kanjilal
Faculty of Lib. & Inf. Sc.
IGNOU, New Delhi

Dr. Neena Talwar Kanungo
Faculty of Lib. & Inf. Sc.
IGNOU, New Delhi

Prof. C.R. Karisiddappa
Head, Dept. of Library & Information Science
Karnataka University
Dharwar

Dr. A. Lahiri
Senior Advisor
Dept. of Scientific & Industrial Research
New Delhi

Prof. P.B. Mangla
Dept. of Lib. & Inf. Sc. (Retd)
University of Delhi
Delhi

Prof. A. Neelameghan
UNESCO Expert (Retd.)
Sarada Ranganathan Endowment for Library & Information Science
Bangalore

Prof. T.N. Rajan
Chief Coordinator (Retd.)
INSDOC, New Delhi

Dr. Pravakar Rath
Faculty of Lib. & Inf. Sc.
IGNOU, New Delhi

Prof. J. Sarkhel
Head, Dept. of Library & Information Science
Vidyasagar University
Midnapur W.B.

Prof. R. Satyanarayana
Faculty of Lib. & Inf. Sc. (Retd.)
IGNOU, New Delhi

Prof. B.K. Sen
Head, (Retd.)
Education Division
INSDOC
New Delhi

Dr. Mahinder Singh
Director
DESIDOC, Delhi

Dr. N. Vijayaditya
Director General
National Informatics Centre
New Delhi

Prof. T. Viswanathan
Ex-Director, INSDOC
New Delhi

Dr. (Ms) Neela Jagannathan (*Special Invitee*)
IGNOU, New Delhi

Programme Coordinators: Prof. S.B. Ghosh and Prof. Uma Kanjilal

Course Coordinator and Editor: Prof. S.B. Ghosh

Course Preparation Team

Unit	Contributor	Academic Assistance
1	Prof. T.N. Rajan	Mr. Anup Kumar Das, Consultant
2	Dr. Jaideep Sharma	
3	Prof. M.P. Satija (Revised by Prof. S.B. Ghosh)	
4	Prof. J. Sarkhel	
5	Prof. J. Sarkhel	
6	Mr. P.S. Mukhopadhyay	
7	Dr. P.K. Panigrahi	
8	Dr. A.R.D. Prasad	
9	Dr. P. Devika	
10	Dr. P. Devika	
11	Dr. P. Devika	
12	Dr. P. Devika	
13	Dr. A.R.D. Prasad	
14	Dr. Aditya Tripathi (Revised by Prof. S.B. Ghosh)	
15	Dr. M.A. Gopinath, Dr. A.R.D. Prasad	
16	Dr. A.R.D. Prasad	
17	Dr. G.G. Chowdhury	
18	Dr. M.A. Gopinath, Dr. J.P. Singh	
19	Dr. G.G. Chowdhury	

Production

Mr. Jitender Sethi	Mr. S.S. Venkatachalam and	**Secretarial Assistance**
IGNOU, New Delhi	Mr. Manjit Singh	Shri S. Swaminathan

June, 2006 (Reprint)

© *Indira Gandhi National Open University, 2005*

ISBN-81-266-2077-3

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 the Indira Gandhi National Open University.

Authors are responsible for the academic content of the courses as far as the copyright issues are concerned.
Further information on the Indira Gandhi National Open University courses may be obtained from the University's Office at Maidan Garhi, New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi, by the Director, School of Social Sciences.

Paper Used : Agrobased Environment Friendly

Laser typeset by : HD Computer Craft, EA 1/75, Inderpuri, New Delhi-110012.

Printed at : Dee Kay Printers, 5/16, Kirti Nagar Indl. Area, New Delhi-110015.

INTRODUCTION

Information is now available in varieties of forms and formats. Due to development of Information Technology it is available not only in print form but also other forms of media such as Microforms, Audio, Video and Multimedia. The complexities of information has also grown over the years. On the other hand the users need and requirement are also varied. To access information from the huge mass and to satisfy the needs of the users it is necessary that the information has to be properly organised and be made searchable so that it can be retrieved according to requirements. This necessitates the development of Information Storage and Retrieval Systems. The ISAR system has gained importance due to the availability of technology in various forms and the facilities it provides.

For the development of ISAR systems the identification of information and its various characteristics is an important factor. The organisation of information requires that the correct meaning and syntax of it should be properly recognised which is an intellectual process. By far this is being done by the human being. Of late, a number of experiments have been conducted and various techniques have been developed for automatic processing and organisation of information. This Course on Information Storage and Retrieval aims at providing you the various aspects of information identification, organisation, storage and retrieval of information.

The Course is divided into following five Blocks to maintain homogeniety of concepts:

Block 1 : Organisation of Information

Block 2 : Bibliographic Description

Block 3 : Content Development

Block 4 : Information Storage and Retrieval Systems

Block 5 : Information Retrieval

Block 1 comprises 5 Units. **Unit 1**, on Intellectual Organisation of Information, discusses the fundamentals of information organisation and their implications in the development of ISAR systems. **Units 2 & 3**, on Indexing Languages, discuss the basic concepts, the problem of information organisation in natural languages, the need for controlled vocabulary and the various types of Indexing Languages. The **Unit 2** also discusses two important types of Indexing Languages – Subject Headings Lists and Thesaurus.The **Unit 3**, on Indexing Languages – Part II, discusses another type of indexing language – Classification Systems with particular reference to Dewey Decimal Classification, Universal Decimal Classification, Colon Classification and others. The **Unit 4** is on Indexing Systems and Techniques wherein the various systems, methods and techniques such as, Pre-coordinate Indexing Systems and Post-coordinate Indexing Systems have been discussed. As vast amount of information is available now on Internet, Web Indexing has also been included in this Unit. The **Unit 5** is on Evaluation of Indexing Systems wherein the criteria and methodologies for evaluation of ISAR systems in general and indexing systems in particular, have been discussed.

Block 2 on Bibliographic Description is concerned with the description of physical items of documentary information. The **Unit 6** describes the Principles and Evolution of Bibliographic Description. The **Units 7 and 8** describe the Rules for Bibliographic Description and the standards available for the Bibliographic Record Format. The **Unit 9** is on the standards in the context of describing Web documents – Metadata.

Block 3 on Content Development is basically in the context of electronic resources content available on the Web. The **Unit 10** describes the norms and guidelines for Content Development. For Content Development, the knowledge of HTML and XML are a pre-requisite which has been discussed in **Unit 11**. The **Units 12 and 13** describe the Web based content development and Multilingual Content development with particular reference to the non-English language content in general, and the Indian languages, in particular.

Block 4 is on Information Storage and Retrieval Systems. In **Unit 14**, the objectives and types of ISAR systems have been discussed. In the context of development of varieties of IR systems, compatibility among the systems will enhance the usability and accessibility of information. These have been discussed in **Unit 15**. A number of intelligent IR Systems have been experienced and developed. The **Unit 16** deals with the various such systems available, so that you are able to understand the Research and Development going on in this area.

Block 5 is on Information Retrieval. In **Unit 17**, the Information Retrieval Processes and Techniques have been covered. Various models of information retrieval have been developed to satisfy the requirements for effective retrieval. These models have been developed taking the inputs from other disciplines. Such models have been described in **Unit 18**. In any IR system, Search is an important activity. Its success and value depend upon this aspect. The **Unit 19**, thus discusses the search strategy and various processes and techniques available in conducting a search effectively particularly in the context of computerised information storage and retrieval systems.

The Course material is in two Parts. Part I (the present volume) covers Units 1 to 9 and Part II comprises Units 10 to 19.

After going through this course you should be able to develop a good understanding and insight into the different aspects of organisation and retrieval of information and also, the development and operation of automated ISAR systems.

BLOCK 1

ORGANISATION OF INFORMATION

ignou
THE PEOPLE'S
UNIVERSITY

UNIT 1 INTELLECTUAL ORGANISATION OF INFORMATION

Structure

1.0 Objectives

1.1 Introduction

1.2 Intellectual Organisation of Information (IOI)

1.2.1 Meaning of Intellectual Organisation of Information (IOI)

1.2.2 Why IOI is Necessary ?

1.3 IOI in Primary Communication

1.4 IOI in Indexing Systems

1.4.1 Derived Indexing Systems

1.4.2 Assigned Indexing Systems

1.5 IOI and Indexing Languages

1.5.1 Classification Systems

1.5.2 Subject Headings Lists and Thesauri

1.6 IOI in User Services

1.6.1 Search Aspects

1.7 IOI and Content Analysis

1.7.1 Meaning and Purpose of Content Analysis

1.8 Information Retrieval Systems – Changing Environment

1.8.1 Search Facilities

1.8.2 Search Strategies

1.9 Examples Illustrating Information Storage and Retrieval Functions

1.10 Future Trends

1.11 Summary

1.12 Answers to Self Check Exercises

1.13 Keywords

1.14 References and Further Reading

1.0 OBJECTIVES

Information and knowledge are basic input to human growth and socio-economic development. The organisation of thoughts and ideas by the generators of information involves intellectual organisation. Over the last centuries, information available in documentary form have grown in volume, variety and forms. To facilitate access to them, secondary tools and techniques are developed involving intellectual organisation of such information and knowledge. This Unit discusses intellectual aspects of organisation – the fundamental aspects and its implications in Information Storage and Retrieval Systems.

After reading this Unit, you will be able to :

- get a clear understanding of intellectual organisation of information;
- distinguish intellectual effort from non-intellectual efforts;
- perceive the intellectual organisation in indexing and indexing languages;

- assimilate the intellectual organisation in user services and searches;
- get an insight into the current trends in computerised information retrieval systems; and
- note current research efforts in information retrieval (IR) systems and their future.

1.1 INTRODUCTION

We have learnt that information and knowledge are prime resources for human growth and socio-economic and technological development of a country. These are deemed to be human capital in the context of economics. A reservoir of knowledge has been accumulated during the course of human history. In the recent century, this resource has grown not only in great volume and variety but also with accelerated speed. This is largely because of development of communication, skill and media that has enabled exchange of ideas among peers with critical evaluation of generated information and knowledge.

We know that information and knowledge are the products of human intellect. Experience gained in the judicious application of information and knowledge to all-round human development, gets added value to this intellectual resource. This is a never ending process – dynamic and ever growing.

Information and knowledge are invariably presented in a variety of forms. Primary information presentations are in the form of journal articles, conference and research papers, technical reports, theses, monographs, treatises and in similar forms. Access to the contents of these forms is provided through indexes to these documents. Primary information when grows rapidly, secondary communication services, in the form of indexes, abstracts and other methods of condensation are developed, which provide access to them.

When we talk of intellectual organisation of information, in our context, it refers to the intellectual methods by which the secondary communication can be organised to provide access to primary communication. In the early 1960s, a series of seminars was organised by the Graduate School of Library Science, Rutgers University, New Jersey, USA, on the Intellectual Organisation of Information with reference to a few systems of classification, indexing systems, techniques and tools, and newly emerging thoughts and ideas at that time. Since then the concept of intellectual organisation of information refers largely in the context of design and development of techniques and tools of information storage and retrieval systems.

If the concept 'intellectual organisation of information' is examined from a broader perspective, almost from the generation and to the final presentation of information, the efforts to organise thoughts and ideas of generators have always been an intellectual effort. At every successive stage, i.e. providing secondary services in the form of storage and retrieval tools and techniques involved in them, have also been the intellectual efforts. For instance, reference interviews with users to provide precise information sought, storage methods, search techniques, construction of depth schedules of classification, design and development of macro and micro thesauri, retrieval efficiency measures, evaluation of indexing systems and the final presentation of retrieval results – in fact everyone of these, indeed, constitute intellectual efforts. With the advent of computers and communication technologies, the scope of searching a variety of databases through Internet and online databases have widened substantially, with several new approaches to information storage and retrieval.

In this unit, we shall discuss what constitutes intellectual organisation of information with a working definition to understand its meaning in our context. We shall also

try to understand what is non-intellectual effort, which may be largely routine and repetitive work, somewhat mechanical without much intellectual effort. Many of this type of work are relegated to computers for speed and accuracy. We shall examine the intellectual and non-intellectual efforts in indexing systems, indexing languages, user services, including some general principles and guidelines that may be useful to organise and search information.

1.2 INTELLECTUAL ORGANISATION OF INFORMATION (IOI)

Information is a product of human intellect and is meant for use by others. Therefore it is disseminated and communicated. Productive transfer of information and absorption of the same by users is dependent on the intellectual organisation of information at every level of its transfer. Primary communication of information, most often in textual form, is logically organised and presented. When this information is properly assimilated by users and combines with their internal information, this organised aggregate of information becomes knowledge. Such knowledge, when used in an appropriate manner and applied judiciously to suitable situations, becomes wisdom. The Table 1.1 depicts this process:

Table 1.1: Stages of Human Thoughts and Ideas

Stages	Mental process	Formalisation	Additional Process
1	Observation	Ideas / Data	Instrument Extra-senses
2	Organisation (Logical relation)	Information	Systematic Principles
3	Conference and Compaction	Subject	Fundamental Principles
4	Learning Assimilation	Knowledge	Perception, Modeling, Process
5	Judgment Correlation Application to Contexts	Wisdom	Sensing the Context Retrieval of Knowledge Appropriate Decision

1.2.1 Meaning of Intellectual Organisation of Information (IOI)

We noted that all efforts to organise information are intellectual. But what exactly is this intellectual effort? Human intellect is the power or faculty of thinking, learning, understanding, assimilating, acquiring and organising mental constructs and recording them in a suitable form for communication. Table 1.1 indicates that from the first stage to the last stage of human thoughts and ideas, there is an intellectual effort. While the intellectual efforts vary at every stage, we are more concerned, in this unit, with the intellectual effort that would provide easy access to primary information that are stored in a variety of documents, irrespective of their contents and physical forms i.e. either print, non-print or digital.

John Ziman, a well-known distinguished physicist, describes how 'information' turns into 'knowledge'[1969]. In the context of scientific research, he observes, "The aim of science is understanding not the accumulation of data and formulae. We need to bite into whole concepts, not swallow them piecemeal. The separate pieces of information in the separate primary papers need to be joined to another, fused with one another, welded into a coherent intellectual machine, which may be used as a whole, whether for material benefit or for further scientific exploration. The work of synthesis is quite important as the analytical process of discovery and experiment." He mentions about three aspects of the organisation of knowledge viz.,

(a) Organisation by creation, (b) Self-organisation and (c) Bibliographic organisation. Indexing and abstracting activities are 'bibliographic organisation' and are 'extraction process' as they extract widely scattered pieces of information from primary sources. Reviewing, surveying, consolidation of literature are 'summation processes' which use the secondary sources like indexing and abstracting services to organise the indexing and abstracting services to provide access to primary information. Therefore, secondary services such as indexes, abstracts, bibliographies are not only extremely important feeders to primary information communication products but are also highly essential.

Self Check Exercise

1) What do you understand by 'intellectual organisation' in general and with particular reference to IR systems?

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

.....

.....

.....

.....

.....

1.2.2 Why IOI Is Necessary?

The sole purpose of intellectual organisation of information of secondary services, like bibliographies, indexes and abstracts, is to provide relevant information to users, within a time frame, with speed, accuracy and ease. This can be explained as follows:

In Figure 1.1, AB is a data file of information (documents). To a specific question, CB of the file is assumed to be relevant and AC is not relevant. When a search for information is made, the file gets divided into four unequal parts as shown below:

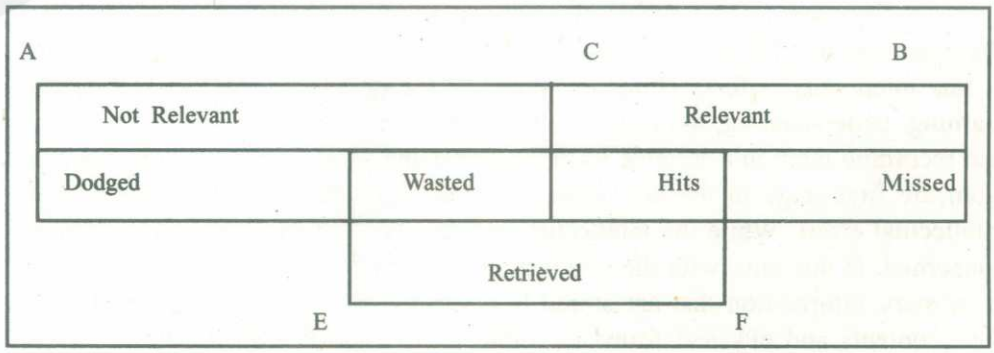


Fig. 1.1 : Aim of intellectual organisation of information

In a search for relevant documents, EF is retrieved, which includes some hits (relevant ones) that are obtained along with some wastes (irrelevant ones) also. The ideal situation would be, if EF and CB exactly coincides, a hundred percent search result would be obtained. But this result is seldom got in practice. So indexing techniques and tools are designed to get as many relevant documents as could be obtained while searching for documents that carry the required information. This applies equally to all computerised IR systems.

The effort in indexing is aimed to get a precise set of documents in a search. The shadow part (Umbra) in Figure 1.2 is surrounded by the neighborhood semi-dark part (Penumbra) and the penumbras are surrounded by unwanted parts (Aliens). So, when a user seeks to use the file of information, the person gets an APUPA pattern facility i.e., he/she gets the precise ones, then somewhat related ones to get a wider range of documents that may carry the required information and successfully avoids the rest that are irrelevant and hence unwanted. To achieve this, IOI plays a most crucial role.

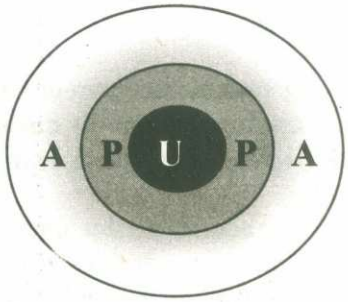


Fig. 1.2: APUPA Pattern
[A=Aliens; P=Penumbra; U=Umbra]

Generally all indexing systems, techniques and tools try to achieve results that could be as ideal as possible.

Self Check Exercise

2) What features are important for accessibility of information?

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 IOI IN PRIMARY COMMUNICATION

Usually macro documents, such as books that include monographs, treatises, project and technical reports, textbooks (particularly advanced level textbooks), do carry indexes to locate the precise place where an idea occurs and is discussed. For instance, a document on 'roses' might carry information with reference to rose cultivation, varieties of roses, rose scent technology, decorative aspects, artificial roses, etc., which are usually scattered in the text of a document. The index to the book collects all these details and provides access to the specific places in which these ideas are discussed in the text. In other words, an index provides a connective facility to the different aspects of a concept, which are invariably scattered in the text. If all the auxiliary words such as prepositions, conjunctions, etc. of a document are taken out of the text, and if only the substantive words or phrases alone are retained, the thoughts and ideas of the author with reference to the topic handled alone will be retained which is the essential part of the document. This set of concepts can be mapped properly to get a full panoramic view of the author's ideas and thoughts. The intellectual skills required here are the ability to understand the subject of the book, choose the substantive words, bringing out the relationship between them and display them in the index in an appropriate manner.

1.4 IOI IN INDEXING SYSTEMS

In every indexing system, indexing language, user services, etc. a part of the indexing work involved will be intellectual and a part will be non-intellectual (clerical as A C Foskett [6] prefers to refer); but both have equal bearing on the index retrieval efficiency. In this Unit, we are merely indicating the intellectual part of the indexing systems and distinguishing them from the non-intellectual ones. In the later units of this course, all the aspects of information retrieval systems are discussed in detail.

A.C. Foskett divides indexing systems into two groups viz. Derived Indexing and Assigned Indexing. It is a useful approach to distinguish their differences. We shall also discuss IOI with reference to these two different types of indexing.

Of all approaches to provide access to documents in indexing systems, the subject approach is the one that posses problems of accurate retrieval efficiency. Other approaches like fact retrieval, or the author and title retrieval, etc. do not pose any serious problem.

1.4.1 Derived Indexing Systems

Derived indexing systems derive their information for indexing from the document. They solely rely on information, which are available in the document without the necessity of adding anything by the indexer and/or from any other source. Derived indexing is used in printed indexes – computerised and on-line systems. These include title-based indexing systems like Catchword, Keyword-in-Context (KWIC), Keyword-out-of-Context (KWOC), Keyword-alongside-Context (KWAC) etc. in which the title is the source for indexing. Citation indexing, although uses the title or the bibliographic reference, has a different philosophy. The details of these indexing techniques are discussed in subsequent Unit 4.

In all these type of indexing, intellectual effort is minimal and therefore are suited for computer operations that permit high speed, efficiency and accuracy.

Expert Systems

Expert system is also a computerised system, which involves a high intellectual capability both in designing and operation. A.C. Foskett summarises the features of and requirements for an expert system as follows:

An expert system

i) must represent the expert's domain-specific knowledge in the way that the expert uses that knowledge;

ii) must incorporate explanation processes and ways of handling uncertainty;

iii) typically pertains to problems that can be symbolically represented; and

iv) is more tolerant of user errors than conventional programs.

In order to achieve these,

i) there must be at least one acknowledged expert in the subject area;

ii) the sources of the expert's expertise are judgment and experience;

iii) the expert must be able and willing to explain his/her knowledge;

iv) the problem must be well bounded;

v) the problem area must have a real consensus; and

vi) test data must be easily available.

Self Check Exercise

3) Why is it that minimum intellectual effort is not required in derived indexing systems. Name some of these systems.

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

.....

.....

.....

.....

.....

1.4.2 Assigned Indexing Systems

Assigned indexing, as the name suggests, is index terms assigned by the indexer to documents for the purpose of storage and retrieval of Information.

For this process, the intellectual skills required are:

- a fairly good knowledge of the subject(s) for which the index provided;
- ability to analyse documents to identify concepts that are to be represented by chosen terms in the right context;
- recognise the sequence in which the concepts are to be related (this refers to the syntactic relationship between concepts); and
- identification of the relationships between the concepts in terms of their broad and narrow relationships with reference to their meaning (this refers to the semantic relationship).

Once these aspects have been properly accomplished, the index entry format can be determined. The rest of the operations can be relegated to Computer Softwares, as they are mechanical, routine and to a large extent repetitive, depending upon the indexing models that are designed.

English language is used for most of our indexing. Therefore the problems of keyword representation are related to English language forms. Multiword concepts and their sequence for citing them in an entry, spelling of words, synonyms and antonyms, homographs, etc. are some of the problems that need attention in indexing.

Examples:

Spelling of word :	American spelling vs. British Spelling (e.g. Center or Centre)
Synonym :	Elevator or Lift
Antonym :	Ascending and Descending
Homograph :	Plant (a living organism) or Plant (a factory)

The syntactic relation refers to the citation order of concepts in index statement.

The semantic relationship is the hierarchical and affinitive or associative relationships of concepts.

The tools, which are used for indexing are indexing languages which include Classification Systems, Subject Headings Lists, Thesauri, Thesaruofacet, and Classaurus.

We shall briefly deal with the indexing languages in the next section. However, you will learn about indexing languages in Unit 2 and indexing systems in Unit 4 in details.

Self Check Exercise

4) What are the intellectual requirements for assigned indexing? What aspects of assigned indexing could be relegated to computers?

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 IOI AND INDEXING LANGUAGES

Index languages are the tools that are used in indexing with syntactic and semantics features built into some of them. Classification systems, Subject Heading Lists, Thesauri, Thesurofacets and such others can be included in indexing languages. Classification schemes are useful for organisation of books on library shelves, displaying an organised form of their arrangement. They are also highly valuable for the study of the ramifications of subjects and as a means of understanding the relationship between concepts. The display of schedules gives us these facilities.

Subject Headings Lists are useful to understand the relationship among concepts to a certain degree, besides their application in indexing.

Information Retrieval Thesauri are based on three types of relationship among concepts viz. Equivalence, Hierarchical and Associative relationships and their display in an alphabetical sequence.

Thesaurofacets combine classificatory principles and provide alphabetical thesauric features in detail. They can be used for shelf organisation of books in libraries as well as information retrieval thesaurus.

We can also include in index languages, natural language indexing and Free indexing languages, which have become quite common with computerised databases.

1.5.1 Classification Systems

Classification systems such as Colon Classification (CC) has an explicit syntactic system and the schedules of classes and its ramifications are organised in a systematic order displaying the hierarchical and associative relationships through a scheme of notation. A set of postulates and principles determine the design of CC. Classification Systems such as CC are generally known as Analytico-Synthetic Systems.

Classification systems in which the syntactic and semantic features are less explicit are the Bibliographic Classification (BC), Dewey Decimal Classification (DDC) and Universal Decimal Classification (UDC) in order of their analytico-synthetic features respectively. The displays of the schedules are organised systematically by a system of notation.

Library of Congress Classification (LCC) has been fully enumerated scheme, the principles of its design being implicit. The schedules are far more extensive than the other classification systems.

We shall learn more about these general classification systems as well as special classification systems in Unit 3 of this course.

1.5.2 Subject Headings Lists and Thesauri

Subject Headings Lists are highly valuable for indexing. The conceptual relationships are indicated in the list and the choice of terms and preference are given. In recent decades they are introducing many thesaural features. The Library of Congress Subject Headings (LCSH) List is one of the best tools for indexing and retrieval. The Sears List is a shorter version useful for smaller libraries. Medical Subject Headings (MESH) and Subject Headings in Engineering (SHE) are some of the other specialised ones highly used. In the Social sciences, the Public Affairs Information Service (PAIS) Subject Headings List is extensively used.

A number of general and specialised information retrieval thesauri have been published since the 1950s. They are highly useful in indexing, for choice of index terms and displaying concept relations.

The Engineering Joint Council (EJC) Thesaurus and the Thesaurus of Engineering and Scientific Terms (TEST) are among the thesauri used for scientific and engineering subjects. In the social sciences, the Educational Resources Information Center (ERIC) Thesaurus is extensively used.

Thesaurofacets combines the features of analytico-sysnthetic classification and information retrieval thesaurus, each of these can be used independently. The best examples are the 'Thesaurofacet of the English Electric Company' and the 'Root Thesaurus' of the British Standards Institution.

The design and construction of classification schemes and Information Retrieval thesauri require high intellectual competence. Besides the technical aspects of construction, a good background of subject knowledge is required for their design and construction. Their use also requires considerable ability to understand their design and construction. The effective use will depend upon constant use of them in indexing and retrieval.

The use of subject headings lists depends very much upon the understanding of the principles of their design and construction. While the use can be learnt with assimilation of the principles and the methods of use, the construction of these is a highly intellectual effort.

Natural Indexing languages are not really a distinct or stable language in their own right, but are rather the natural or ordinary language of the document being indexed. Here all terms are taken from the document itself.

A free indexing language cannot be listed. Indexing is free in the sense that there are no constraints on the terms that can be used in the indexing process. Free indexing is different from natural language indexing in that the latter is constrained by the language of the document being indexed, whereas the former is not, i.e. the appropriate terms according to necessity may be assigned.

Self Check Exercises

5) What are the types of indexing languages that are usually used in IR Systems?

6) What are the intellectual components in constructing indexing languages?

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

.....

.....

.....

.....

.....

.....

1.6 IOI IN USER SERVICES

The primary purpose of all this intellectual organisation of information (IOI) and knowledge is to serve the various categories of users effectively to satisfy them in their information needs. But the effectiveness of user services depends upon our understanding of users and their information needs.

Information seekers are often in a state of uncertainty or lack of clarity in their conception of their information requirements. Their expression of needs is most often not precise and they need some intellectual support in the form of some hints or cues. To prepare an effective set of cues, it is necessary to analyse their behaviour and style of searching, even if they do their searches themselves or seek the help of information officers. Interaction with users to ascertain correctly their information needs would necessitate formal or informal meetings with them. In this process, some of their problems may get sorted out such as:

- How much technical knowledge do they already have about the subject?
- What position do they hold in their organisation?
- What are their attitudes about the subject or about their nature of work or about searching information?
- How do they use retrieved document(s)?
- How much do they trust or have faith in information professionals assisting them in their quest for information?

1.6.1 Search Aspects

In a library, searches may be carried out by users themselves with the tools such as the library catalogue or printed indexes or computer files, if there are such facilities. In such cases, users can modify the searches as they progress in their search. If they are familiar with search procedures, they may succeed in getting right results; sometimes they may also get exposed to more than what they had actually expected because the nature of the organisation of the files may be versatile.

All users, irrespective of their knowledge level, have a specific reason for using a specialised document, and they need information tailored to their level of knowledge. The difference in need of the required document may be based on their requirement. For instance a nephrologist may require a report on kidney transplantation, seeking high tech information on surgical research. Another specialist in nephrology may be interested in obtaining information to recommend a new surgical appliance for a clinic as a semi-expert who must decide whether

or not to purchase the appliance. A person may be interested in space travel as a non-expert who enjoys learning about space. Another seeker may seek an owner's manual for a new camera as a consumer who is looking for instructions for operating the camera. A software designer may be looking for a document as a generalist to give specification to interact with a computer system analyst.

Thus, we find a multipronged approach to searching information to cater to the needs of different categories of information seekers and give them their required information at par their level of interests.

If the files of information include different categories of physical documents, the variation in search and perusal pattern would lead to a kaleidoscopic analysis, organisation and retrieval. Therefore, the basic issues in intellectual organisation of information concerns the following aspects:

- Analysis of searchers' needs according to their level of technical knowledge, positions they hold in their organisation, attitudes, use of retrieved documents, etc;
- Searchers generally have one of the three levels of knowledge viz. expert, semi-expert and non-expert;
- A searcher's response to information searching and access may be influenced by emotions, motivation and preferences;
- Searchers may differ in their information requirements. Some may search only for an abstract or an introduction to the topic; others may look for specific sections or topics in a document; still others may want the whole text search; and some would like a review or a critical evaluation of the information available on the field. Some may need hard copies of the documents;
- When a set of documents has multiple searchers, there is a need to develop content organisation and display to serve both primary and secondary information searches;
- Information searching and organisation has to be tuned to the needs of the searchers. Interaction and discussion with them with reference to their information needs would be useful; and
- When one type of intellectual organisation of information cannot meet all the needs of searchers and cannot provide a multiple display, it is necessary to provide multiple organisations and displays, so that the searchers can approach the information system from several points of view.

Self Check Exercise

7) Discuss briefly the purpose of information service provider interacting with information seeker, indicating the intellectual components and others.

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

.....

.....

.....

.....

.....

1.7 IOI AND CONTENT ANALYSIS

For intellectual organisation of information, one has to first analyse the content of literature. We know that *contents* usually refer to thought and ideas, which a person wants to communicate. In the context of bibliographic information, the

contents are the thoughts or ideas, which an author wants to communicate to others with the expectation that these will be used by someone for varieties of requirements. These may be descriptive, explanatory, investigative examining the pros and cons of a particular set of ideas or a phenomenon, experimental or imaginative or creative. The physical media through which contents are communicated may not be paper-print alone but also a variety of non-print media.

To facilitate the use of contents by users according to their varied needs and requirements, it is necessary that the contents are properly analysed, essential concepts are identified which are organised in a manner to be usable by the users. Thus, content analysis assumes a great importance for organisation of information sources and providing varieties of information services and products.

1.7.1 Meaning and Purpose of Content Analysis

Meaning

Some of the formal definitions of content analysis are given below to enable you to have fuller comprehension of the meaning and scope of it.

Content Analysis is the effort of analysing the records of human experience and of knowledge; it is a means of studying all types of communication, its nature, its underlying meanings, its dynamic processes, and the people engaged in the act of communication.

Content Analysis is concerned with the message phase of a three-step communication process, viz., (i) a sender's various motives to produce a message that is intended to express these motives for a receiver, (ii) the effect of the messages expected upon the designated audience, and (iii) on the relationship between messages and its antecedents or the senders producing them.

Content analysis is obviously an activity that follows the creation and production of a document, or a speech, or any observation of a phenomena. The purpose is to infer or interpret what the message are about and what their effect will be on the users. More precisely, in the field of library and information services, content analysis is a means to organise the intellectual content or documents, irrespective of their physical form, so that the users get easy access to their source of information.

Purpose

The purpose of content analysis varies depending upon the activity to which this technique is applied. In library and information Science, content analysis lies at the root of many of the operations and services of libraries and information centres. With the exponential growth of recorded communication in a variety of forms and languages, three significant events have occurred. These are:

i) information storage and retrieval has assumed great importance and has become a highly specialised subject in which persons from different fields are increasingly getting involved;

ii) content analysis has developed as a research method that complements other methods of investigation; and

iii) the man-machine interface has developed user friendly systems of databases to explore the users intentions and meanings and to analyse the effect of resources that has intensified.

Self Check Exercises

8) What is the meaning of 'Content Analysis'?

9) Explain the purpose and use of Content Analysis.

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

.....

.....

.....

.....

.....

.....

1.8 INFORMATION RETRIEVAL SYSTEMS – CHANGING ENVIRONMENT

We have so far been discussing some general aspects of information storage and retrieval systems, their purpose and functions, their intellectual and non-intellectual components, users approach to information etc. This section attempts to deal with a few of the current trends in information retrieval systems and makes a summation of them. The detailed descriptions of these IR systems will be made and discussed in the units specifically devoted to them (Unit 13 to Unit 18).

The three features of current information retrieval systems that are very conspicuously distinct are:

- IR systems are increasingly computerised, although printed indexes and card catalogues in libraries still exist. But these are also switching over to computerised systems and will get fully automated in the near future.
- Computer databases have grown phenomenally and are in all conceivable areas. Global databases/ networks in academic disciplines that include all education and research materials in almost every specialised areas, business and industry, government resources, current affairs, finance, legal matters are all burgeoning rapidly. These are accessible through Internet and on-line services through commercial information service vendors. Full texts of documents, document surrogates like indexes, abstracts, and a host of bibliographic and other forms of information are accessible to users. Digital representations of information and multimedia technologies have given further versatility to IR systems.
- Users also have grown with a large number of them making direct access to these databases in a variety of subjects. These users are domain experts (with or without any knowledge of using computerised databases), laypersons and information intermediaries (who have gone through systematic professional training and experience).

The key issue is that any retrieval process is very much dependent upon the indexing and storage stages. In fact these two intellectual activities determine the search strategy to a large extent to obtain optimum results. However, the nature of computerised indexes varies greatly and searchers will get successful results if they recognise some of their inherent strengths and limitations. Generally speaking however, there a number of features that are common to all information retrieval systems, including CD-ROM based systems, the Internet, Online search services, Document management systems, and OPACs. There is a trend in using all these information networks or media by all kinds of users that include novices,

domain experts who constantly use IR systems and are familiar with their facilities and functions, occasional users who may depend upon information intermediaries, and other type users who have special needs. Information managers and intermediaries, who do searches on behalf of all categories of users, are expected to be expert users of IR systems. They must work with the users, understand their specific needs and do the search in collaboration with users, reviewing the results of retrieval at every search stage to get maximum results.

Self Check Exercise

10) What are the distinct current trends in IR systems?

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

.....

.....

.....

.....

.....

1.8.1 Search Facilities

In most computerised information retrieval systems, standard search facilities have been developed. These facilities provide access to full texts, and through pre-assigned keyword words. Some of these facilities are listed in Table 1.2:

Table 1.2: General Facilities Available in Computerised Information Retrieval Systems

Facility	Purpose	Whether Intellectual Activity
Set-up, e.g. Help	Set up the environment	No
Searching search terms	To identify search terms viewing index terms from standard lists such SHL or thesauri	Yes
Entering search terms	Searcher to entry terms	No
Combining search terms	To develop search strategies through search logic	Yes
Searching fields	To make choice of fields in which terms(s) appear	Yes
Truncation	Text strings using truncations	No
Syntactic/semantic	To set words in combination and semantic relations	Yes
Setting specific range	To set ranges, e.g. numeric publication date	No
Displaying search results	To Show no, of references	No
Displaying records	To display records on the screen	No
Search management	To review search	Yes
Display the thesaurus	To display syntactic/semantic relations	No
Hyperlinks	To navigate for associated terms	Yes

The facilities with their purpose and nature (Intellectual or non-intellectual) are briefly explained below:

- Set-up facilities: These are normal system facilities to operate a database.
- Selecting terms: This involves selection of search terms that can be obtained through display of search-terms. This is an intellectual activity.
- Entering terms: A routine operation
- Combining search terms: Search terms may be combined into search statements with the aid of search logic. This is an important IOI activity to get proper results.
- Searching fields: This is a facility to search for the occurrence of terms in

a specific section of a document or in specific fields in a record to get more precise searching. This is an important IOI activity.

- Truncation: This is a normal routine activity to facilitate wider retrieval.
- Syntactic/semantic relations: This is an intellectual activity to set the search terms in context and also to widen search results.
- Setting specific range: This is a routine activity.

The next two facilities are routine activities displaying the results

- Search management: This facility permits reviewing search results and constitute the most vital results of the search.
- Hyperlink: Hyperlink is an excellent facility that permits navigating search files to get associated terms available in different sections.

1.8.2 Search Strategies

The facilities built into computerised information systems, briefly stated above, offer to conduct effective and successful results. The set of decisions, to be taken at different stages, constitutes the search strategy. The searches ought to be methodical and the construction of search strategies should yield best results. Often it will be necessary to broaden or narrow down a search dependent on the search result in the first attempt. Iterative searches or heuristic searches will permit search result to be reviewed at every search and modified the search strategies as the search progresses to get maximum successful results.

What have been stated above are some broad indications of accessing databases for getting right information or references in a computerised information systems.

1.9 EXAMPLES ILLUSTRATING INFORMATION STORAGE AND RETRIEVAL FUNCTIONS

In this section, a few examples are presented exemplifying indexing, storage, search and retrieval of information. Examples illustrate the intellectual and non-intellectual efforts involved in these functions. Examples also represent typical queries of different categories of users and in different subjects with their contexts. Recall and precision measures, the concepts of exhaustivity and specificity are also explained while discussing the searching and retrieval aspects. But these are described just in brief.

Example 1

A textile research scientist is looking for research papers on the chemical treatment of cotton fibers for fire proof resistance in the production of garments for men.

Example 2

An anxious young mother is interested in information on the problem of dyslexia of Indian children at primary and secondary levels of education in learning English

Example 3

A student trying to get admission to MBA course, wants to know the curriculum details on the courses on Personality development and on the Growth and Progress of Indian Socio-Economic and Cultural studies respectively of the Indian Institute of Public Management, located at Delhi in 2004.

Example 4

A storywriter for children wants to know if there is a equivalent of Cinderella story in Indian folk tales.

We shall discuss all the operations involved with reference to (i) Indexing, (ii) Storing the index words in a file (iii) Searching a file and finally (iv) Retrieving.

As a perquisite for all these efforts, it is necessary to point out that the information professional should have a fairly good knowledge of the subjects involved, ability to read fast and understand a document(s) to identify the main thought contents of the documents.

Presuming that an information professional tries to provide the service, we shall discuss the various operations involved viz. Indexing, Storing, Searching and Retrieving.

Steps in Indexing and Storing

- After identifying the main thought contents of a document(s), a subject statement comprising appropriate keywords has to be formulated.
- The keywords representing the different facet ideas of the document have to be stated according to the syntactical rules of the indexing system chosen, particularly if an assigned indexing system is chosen.
- The indexing terms could also be stated without their contexts at the time of indexing. In computerised systems, the keywords could be input using a RDBMS (Relational Database Management System) Software. At the time of search, the keywords could be combined for retrieval using Boolean operators.
- To provide additional references for getting all the relevant documents to the query in a search, the semantic relationships of the keywords have to be taken into account and have to be rendered, using a standard authority list such as thesauri, Subject Headings Lists, Classification schedules or such others.

These four operations are the most important intellectual part of indexing.

Once the keywords representing the thought contents of the documents are formulated and rendered in the standard form(s) displaying syntactical relations, the rotation of the formulated subject string can be manipulated by a software program to be input into the computer file. These operations are mechanical, repetitive and no intellectual effort is needed.

Steps in Searching and Retrieval

- In formulating a search statement for retrieval, the information professional should have a clear idea of what exactly the user is looking for. An interview with the user must be the first step in formulating a search statement. It may be necessary to do iterative searching getting appropriate feedback and response from the user for further searching.
- Formulating the search statements with appropriate standard keywords in their right spelling.
- Searching the file for obtaining the right documents.
- Scanning the results, choosing the right references and getting the hard copies of the document(s) if necessary.

Now, we shall discuss each of the examples given above and apply these principles and procedures.

Example 1

Process of Indexing

The formulation of the subject statement represented by keywords, assuming the subject approach of the majority of users, is as follows:

Textile technology, Cotton fibers, Fire-proof resistance, strengthening, chemical treatment, men garments

The keywords are cited according the postulates and principles of syntactical rules associated with an assigned indexing system

Keywords in computerised systems can be entered in their right spelling, using a RDBMS Software.

Additional references could be provided by finding the associated, hierarchical and equivalence keywords using a thesauri for Cotton fibers, Fire proof resistance, chemicals for strengthening, men garments. The associated terms for each of the keywords are given below:

Cotton fabrics: Woolen fibers, Synthetic fibers

Fire proof resistance: Fire intensity, Optimum temperature

Strengthening: Reinforcing

Chemicals: Name(s) of Chemicals

Men garments: Shirts, Slacks, T Shirts

Additional **See also** references may be provided for getting a fairly good retrieval of relevant documents.

These are activities combining a fairly good knowledge of textile technology with all its ramifications and professional indexing skills and stating them in standard forms. The rest of the operations for storing the keywords in a file and providing additional entries, etc. may be turned over to computer operations.

Searching and Retrieval

In the formulation search statements, the steps starts with a dialogue with the user to ascertain his/her specific need and other information as stated in the earlier section.

In this example, if the subject statement comprises, search words that match the keywords indexed, the search results could be reasonably good.

Boolean operators can be used in searching to state the syntactical relations.

The rest of the operations can be executed by the computer or manually without much of intellectual effort.

Example 2

Indexing and Storing

In this example, the word Dyslexia has to be understood. This expression refers to the abnormal difficulties of children in spelling and reading a language. This is particularly widespread in India today. The young lady wants comprehensive information on getting the different kinds of difficulties in learning a language as well as methods that can correct the deficiencies.

The formulation of the subject statement comprising keywords may be as follows:

Education, Indian Children, Language Learning, Dyslexia, Corrective methods
Additional references could be:

Indian Children: Different linguistic groups

Language learning: English, Hindi, and other Indian languages

Dyslexia: Learning disabilities, Slow learning, Memory

Corrective methods: Teaching techniques, Specialised counseling

Searching and Retrieval

The same steps could be used as shown in *Example 1.*

Example 3

In this case, the student wants accurate information about MBA course details on Personality Development and Current socio-economic-cultural studies of India, as he/she prepares to do the course. Personality development has wide dimensions that include behavior patterns, public relations, and communication. Both written and oral, and a number of other skills. Socio-economic and cultural relations have also wide dimensions that include state policies, fiscal measures, social and regional imbalances, cultural heritage and similar others. The student, apart from getting information about these courses in institution's syllabus, might be interested in getting to know more about these subjects. Therefore handling this query requires not only searching databases and Websites of IIPM but also contacts with teachers and past students of the course.

The formulation of the search statement comprising keywords may be as follows:

Education, Post graduate professional courses, MBA, Curriculum, Personality Development, Indian Socio-Economic-Cultural Studies, IIPM.

Additional references:

MBA Curriculum: Personality Development, Public Relations, Communication Skills

Indian socio-economic-cultural studies: State policies, Fiscal Measures, Budget, Growth Rate, Cultural heritage

IIPM: IIM, Business Schools, Management Institutions

The sources from which information have to be obtained may be varied including personal contacts.

Example 4

This example is different from all the above three examples. The answer to this query may net be got through the bibliographical database searches. Anthologies of folk/fairy tales of India by region and language may be required.

(Cinderella, a heroine of a fairy tale who is oppressed by a malevolent stepmother but achieves happiness through the intervention of a fairy grandmother)

The search statement may be as follows:

Folk tales: Anthologies of folk tales, English translations, Indian language, Regional folk tales, Indian fairy tales, Cinderella story

Additional references:

Literature; Fiction, Short Stories

Anthologies of folk tales, Fairy tales

Indian language: Folk tales

Cinderella stories, Translations

Assuming that in all the four examples, we get a hit relevance of 40, we can work out the Recall ratio as follows:

Again assuming that the relevant references in the database is 100, we get a relevance ratio of 40/100 i.e. 40 per cent

It must be noted here that the total relevant references for a particular query is not generally obtainable in a file, unless every reference in the file is scanned and noted for its relevance to the query. Hence, most often the relevance number is assumed. To this extent, the measure of recall will always be subject to question. Yet it is generally used as a measure for search performances.

The precision ratio would then work out to be, assuming we get a precision of 5 references, i.e., $5/40 \times 100 = 12.5$ per cent

We can expand the relevance by using more and more broader terms in the search, using the principle of Exhaustivity.

Similarly we can get precision by introducing more and more narrow terms and fixing them in context, using the principle of Precision.

Further explanation on Recall and Precision aspects are provided in Units 4 and 5.

1.10 FUTURE TRENDS

Active research efforts are pursued to refine or develop efficient information retrieval systems all over the world. The primary purpose in all these efforts is to get the best results for users. System developers and suppliers are striving to offer a best software package or a service. The information researcher is also equally concerned to develop the best retrieval system. The directions, these research activities are taking, may be summed up as given below:

Better Systems Design

This is primarily to improve methods of matching document descriptors with query descriptors. New methods of searching, other than Boolean search logic has been a major area of further research. Optimum retrieval efficiency, providing a reasonable proportion of relevant hits, and reducing overload of irrelevant hits is another line of research. Speed of retrieval is also implied in this type of research efforts.

Improved Retrieval Facilities and Strategies

Here an attempt is made to improve the efficiency and effectiveness of the system, including the characteristics of the storage requirements, the retrieval speed and the effectiveness of the system. Work in this area seeks to overcome the limitations of inverted file (a term entry system as against an item entry system) by developing fast methods of scanning the contents of database. This involves improving the speed of searching by a set of text scanning algorithms; alternatively to seek hardware-based solutions, most of which are related to speed of text-scanning.

Human-Computer Interface

This is another area of intense research. User-friendly approach is focused on development of self-explanatory, intermediary computer systems, that would stimulate best-match searching, using knowledge-based techniques taking cue from artificial intelligence.

A very new research trend is surfacing, beginning from 1991, a sort of reawakening of Ranganathan's ideas of postulates and principles for designing of faceted classification systems. Quoting from Parthasarathy [2004]:

"Persons engaging in designing tools and systems are known as Information Architects and their profession as Information Architecture. The contributions of Ranganathan, his Laws, Principles, Categories, Canons, Postulates, etc. are now appearing on the Net freely, with occasional comments or explanations, for the benefit of Information Architects. During the last ten years, papers have started coming on the relevance of Ranganathan's ideas for improving the Net and Web-based services. Classification schemes are used on the Net. They are generally enumerative schemes. It is stated that Colon Classification, which is a faceted scheme with analytico-synthetic approach, will be ideal for improving web-based searches. We should take advantage of this opportunity and develop tools and systems for web-based services on Ranganathan's ideas with the help of software groups in India."

References on the work of the Information Architects is appended to the list of references given at the end this Unit [Parthasarathy, 2004].

While it is difficult to foresee to what extent these research efforts would give desired results, the pursuit of continuous research is essential. Domain experts, information specialists, software and hardware experts constitute the research team in these areas of work. Many of these efforts are basically intellectual. Once the design structure gets stabilised, the operational part could be taken care of the technologies.

Self Check Exercise

11) State the areas of current research in IR systems.

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

.....

.....

.....

.....

.....

1.11 SUMMARY

This unit identifies the intellectual and non-intellectual activities that go to make an efficient information retrieval system. In the rapidly growing context of information and knowledge, the efforts to retrieve relevant information from a variety of information networks become absolutely essential. This various sections of this unit highlight these aspects, indicating measures of retrieval efficiency. All these are looked in the context of computerised systems, information networks and users' approach. Research efforts are continuously made to get the maximum efficiency of IR systems.

1.12 ANSWERS TO SELF CHECK EXERCISES

1) Human Intellect is the power or faculty of thinking, learning, understanding, assimilating, acquiring and organising mental constructs and recording it for communication.

In the context of IR, intellectual organisation refers to the intellectual effort to organise the storage and retrieval of information and knowledge, serving

users to access their required information precisely, accurately and with speed and ease.

2) To provide a facility for accessing primary information, secondary services are developed. It should be done in a manner that should provide ease of access, with speed and accuracy and should precisely meet the information requirements of users. Without this effort, accessibility to valuable information will be more or less lost.

3) Derived indexing systems derive their information from the document(s) for indexing and are therefore highly amenable for computer organisation with least intellectual effort. KWIC, KWAC, Citation indexing are a few examples.

4) Assigned indexing require intellectual ability to assign keywords to the thought contents of documents in such a way that these selected terms could effectively represent the contents and make it possible to retrieve the precise information sought by a user. It also involves effective use of framing multi worded search statements, fixing the context. For this the indexer should have a good knowledge of syntactic and semantics of indexing. Assigned indexing require intellectual ability to assign keywords to the thought contents of documents in such a way that these selected terms could effectively represent the contents and make it possible to retrieve the precise information sought by a user. It also involves effective use of framing multi worded search statements, fixing the context. For this the indexer should have a good knowledge of syntactics and semantics of indexing languages.

   The routine operations of arranging the references in a suitable format and order and matching the search terms with the text of documents or surrogates according to inbuilt specific rules can be handled by computer easily.

5) Classification systems, Subject Headings Lists, Thesauri, Thesaurofacets are some of the types of indexing language.

6) The intellectual components in construction of these indexing languages, are the selection of terms, establishing their relationships and indicating them, display of these terms, etc.

7) Interacting with an information seeker, an information service provider, should try to extract as much information as possible to get the needs of the user. This is to formulate a query statement, which should help in deciding on the search strategy. The search result depends on the knowledge the information provider has on the database that is being searched and the ability to get right results, consulting the user. More than one search may often be required to get to the precise results.

   The intellectual component in this effort is the ability of the information provider in formulating the search strategy, after eliciting the right response from the information seeker.

8) Content analysis is the effect of analysing the record of human experience and of knowledge; it is a means of studying all types of communication, its nature, its underlying meanings, its dynamic processes, and the people engaged in the act of communication.

9) The purpose of content analysis varies depending upon the activity to which this technique is applied. It is used as a research technique in social sciences, in psychology for counselling and diagnostic studies in psychiatrics. In library and information science, content analysis is used for a number of activities such as indexing and abstracting, classification of user studies, and in the production of a number of information products and user services.

   It can be used for varieties of purposes, for example, in the analysis of voting pattern in elections. Analytical newspaper reports, studies by institutions who specialise in election predictions, election manifestoes and other types of

analysis often made by experts in the field available in the form of published documents, government reports, academic publications, etc. The analysis of all these, will give a comprehensive view of the different shades of opinion and facts, which will lead to a new set of interpretations and implications. Content analysis form the basis of such studies.

10) IR systems are increasingly becoming computerised, although printed indexes and card catalogues in libraries still exist. But these are also switching over to computerised systems and will get fully automated in the near future.

Computer databases have grown phenomenally and are in all conceivable areas. Global databases networks in academic disciplines that includes all education and research materials in almost every specialised areas, business and industry, government resources, current affairs, finance, legal matters are all burgeoning rapidly. These are accessible through Internet and on-line services through commercial information service vendors. Full texts of documents, document surrogates like indexes, abstracts, and a host of bibliographic and other forms of information are accessible to users.

Users also have grown with a large number of them making direct access to these databases in a variety of subjects. These users are domain experts with or without any knowledge of using computer-databases, laypersons and information intermediaries who have gone through systematic professional training and experience.

11) Better Systems Design; Improved Retrieval Facilities and Strategies, and Human-Computer Interface.

1.13 KEYWORDS

Algorithm : Instructions for carrying out a series of logical procedural steps in a specific order.

Boolean Search : Searching databases using Boolean operators, like, AND, OR, SAME, NOT, in same search field or a combination of search fields.

Exact Match : Searching databases on the topic of a phrase (a combination of words) entered by the searcher that exactly matches.

Best Match : Searching databases on the criteria given by the searcher that almost matches.

Exhaustivity : Navigating in a database, choosing broader search terms to obtain greater number of references.

Heuristic Search : User's search that gets modified continuously in the course search, in the light of knowledge gained at each stage of search results.

Intellectual organisation of information : The power or faculty of the intellect and the mind of thinking, learning, understanding, assimilating, acquiring and organising mental constructs and recording them in a suitable form for communication.

Iterative search : Searches of information service provider that get modified in the light of the interaction with users reaction to the search results.

Kaleidoscopic analysis : Pertaining to changing, complex, variety of forms/images in a manner suggesting changing pattern of a kaleidoscope. A kaleidoscope is an optical instrument in which bits of glass, beads, etc. hold loosely at the end of a rotating tube are

shown in continuously changing symmetrical forms by reflection in two or more mirrors set at angles to each other.

Keyword Search : Searching databases on the topic of a keyword entered by the searcher.

Penumbra : Partly shaded area around the shadow of an opaque object, referring the related areas of information to get additional information relevant to a query.

Semantic : Meaning of words and their relationships

Specificity : Obtaining exact references to query, choosing very specific terms in a search.

Surrogate : Person or thing that acts or is used instead of another as substitute; an entry standing for a document.

Syntactic : Words and their order of representation in a multiword entry of an index file.

Umbra : Dark central part of the shadow cast by the earth or the moon in an eclipse, referring here to the exact references to a query.

1.14 REFERENCES AND FURTHER READING

Bates, Mike (2002). After the dot bomb: getting the web information retrieval right this time. *First Monday*, 7(7), 1-8.

Clarke, Dave and Yancey, Trish (2001). *Twenty-first century tools for vocabulary management and indexing*. Presented at the annual meeting of the American Society for Information and Technology, Washington D.C.

Denton, William (2003). *How to make faceted classification and put it on the web*. <http:/www/miscatonic.org./Library/facet-web-howto.html>

Denton, William (2003). *Putting facets on the web: an annotated bibliography*. <http:/www/miscatonic.org/ Library/facet-biblio.html>

Ellis, David and Casconseleos, Ana (1999). Ranganathan and the net: using facet analysis to search and organise the World Wide Web. *Aslib Proceedings*, 51(1), 3-10.

Foskett, A. C. (1996). *The subject approach to information*. 5th ed. London: Library Association Publishing.

Glassel, Aimee (1998). Was Ranganathan a Yahoo! *InterNIC News*, 3, 1-4.

Godert, W (1991). Facet classification in online retrieval. *International Classification*, 18, 98 – 103.

Gopinath, M.A. (2002). *Intellectual organisation of information*. In: MLIS –3, Block 1, Unit 1. New Delhi: IGNOU.

Ingwerson, Peter and Wormell, Irene (1992). Ranganathan in the perspective of advanced information retrieval. *Libri*, 42(3), 184-201.

Lancaster, F.W. (1991). *Indexing and abstracting in theory and practice*. London: Library Association Publishing.

Louie, Aaron J; Maddox, Erie L and Washington, William (2003). *Using faceted classification to provide structure for information architecture*. 2003 IA Summit.

Marchionini, Gary (1998). *Information seeking in electronic environments*. Cambridge: Cambridge University Press.

Neelameghan, A. SR Ranganathan's impact on knowledge organisation tools. *Information Studies*, 6(2), 77-80.

Parthasarathy, S. (2004). *References to the work of the information architects on applying Ranganathan's ideas*. In: *Summary of the Sixth Professor S Dasgupta Memorial Lecture* delivered on March 10th, 2004.

Ranganathan, S. R. (1968). *Prolegomena to library classification*. 3rd ed. Bangalore: Sarada Ranganathan Endowment of Library Science.

Ranganathan, S. R. (1989). *Elements of library classification*. 2nd ed. Bangalore: Sarada Ranganathan Endowment of Library Science.

Rowley, Jennifer (1998). *The electronic library*. 4th edition of computers for libraries. London: Library Association Publishing.

Spiteri, Louise (1998). A simplified model for facet analysis. *Canadian Journal of Information and Library Science*, 23, 1-30.

Steckel, Mike (2002). Ranganathan for IAS: An introduction to the thought of Ranganathan for information architects. *Boxes and Arrows*, Oct. 7, 1-4.

Vickery, B.C. (1970). *Techniques of information retrieval*. London: Butterworth.

Ziman, John (1969). *Public knowledge; the social dimensions of science*. Cambridge: Cambridge University Press.

ignou
THE PEOPLE'S
UNIVERSITY

UNIT 2 INDEXING LANGUAGES – PART I : CONCEPTS AND TYPES, SUBJECT HEADINGS LISTS AND THESAURI

Structure

2.0 Objectives

2.1 Introduction

2.2 Indexing and its Types

- 2.2.1 What is Indexing
- 2.2.2 Indexing and Classification
- 2.2.3 Types of Indexing

2.3 Indexing Language

- 2.3.1 Definition
- 2.3.2 Need and Purpose
- 2.3.3 Characteristics
- 2.3.4 Types

2.4 Vocabulary Control

- 2.4.1 Controlled Vs. Natural Language Indexing
- 2.4.2 Methods of Vocabulary Control

2.5 Classification Schemes

2.6 Subject Headings Lists

2.7 Thesaurus

- 2.7.1 Definition
- 2.7.2 Functions
- 2.7.3 Construction

2.8 Thesaurofacet

2.9 Classaurus

2.10 Sears List of Subject Headings

2.11 Library of Congress List of Subject Headings

2.12 Summary

2.13 Answers to Self Check Exercises

2.14 Keywords

2.15 References and Further Reading

2.0 OBJECTIVES

Retrieval of information by subjects from huge mass of documents requires that essential concepts are identified and organised in a searchable form. Indexing is a mechanism by which information contained in documents can be organised. Problems lie with identifying and organising the concepts. In the documentary information, authors communicate in natural languages which are characterized by linguistic features. To overcome the problems of natural language, the need

ignou THE PEOPLE'S UNIVERSITY

for an artificial language – indexing languages arises. In this Unit we will be discussing the concept of indexing language and its types and also two specific types of indexing language – Subject Headings Lists and Thesauri.

After reading this Unit, you will be able to:

- understand the concepts of indexing and indexing language;
- get acquainted with different types of vocabulary control devices;
- appreciate the differences between a list of subject headings and thesaurus;
- know in depth the provisions of Sears List of Subject Headings and Library of Congress (LC) List of Subject Headings; and
- get an insight into the provisions in a thesaurus and methodology of its construction with reference to application of computers.

2.1 INTRODUCTION

Libraries and librarians strive to acquire and make available the information that exists to satisfy the needs of all concerned. While doing so they take care that the users of information come across its existence and lay hands on it easily. A number of tools and techniques have been developed over the years and indexing is one such technique. It is a technique by which information available in the documents is represented and organised to enable easy access and retrieval. Of course, classification and cataloguing perform the similar function. Then the question arises as to why indexing is necessary and how it is different from other techniques? The difference arises due to the purpose and the types of information for retrieval which decides the functions to be performed. The purpose of classification and cataloguing is basically to organize and provide access to macro information, whereas indexing aims at providing access to micro information. Function of classification is to enable the users to browse the documents on shelves or in a catalogue whereas that of indexing is to enable access to information contained in the document/literature through subjects. Subject cataloguing more or less performs the same function.

In the previous Unit you have learnt about the importance of intellectual organisation of documentary information. Whenever you create a subject surrogate, intellectual organisation plays a crucial role. Indexing is a mechanism of creation of such a surrogate file. This Unit discusses the fundamentals of indexing, the problems encountered in indexing, and how to overcome them to facilitate efficient retrieval of information.

2.2 INDEXING AND ITS TYPES

2.2.1 What is Indexing?

The word 'Index' comes from the Latin word 'indicaire', meaning 'to point out or to guide'. The art or technique to prepare such guides is indexing. According to British Standards (BS 3700:1964), index is "a systematic guide to the text of any reading matter or to the contents of other collected documentary material, comprising a series of entries, with headings arranged in alphabetical or other chosen order and with references to show where each item indexed is located". Indexing is, thus, a process by which the information is organized to enable its easy retrieval and access. Subject indexing "refers to the process of identifying and assigning labels, descriptors, or subject headings to an item so that its subject contents are known and the index, thus created, can help in retrieving specific items of information." [Unesco Handbook of Information Systems and Services,1977].

2.2.2 Indexing and Classification

The purpose of indexing and classification are grouping and, thus, have resemblance to each other. Both refer to processes that involve analysing the subject of the document to represent and organising them for easy access later. We make use of different tools and techniques in the two processes. Indexing makes use of an indexing language to represent the concepts and classification makes use of a classificatory language. The result of indexing and classification is also different. Indexing results in an index whereas classification results in a class number.

Index is a verbal representation of the subject contents of a document whereas the class number is represented in numbers or any other may be having ordinal value. Index provides access to information in an ISAR system through various surrogate of the documents. Class number helps to arrange the documents on shelves according to their subjects. The arrangement of the documents on shelves is in a near neighbourhood relation. Documents on closely related subjects are brought together. Ranganathan called this as APUPA arrangement. It helps the searcher to have a panoramic view of the documents and, thus, browse while searching for his documents. Similar display of document surrogates would not have been possible by verbal representation of subjects. To enable such display, indexing languages make use of different techniques.

2.2.3 Types of Indexing

Indexing is of two types, viz, derived and assigned. Derived indexing uses the same language as that used by the author. It is also known as Natural Language Indexing (NLI). Words/Terms used by the author in the text are used to provide access to users. Such a system of indexing suffers from a drawback which is, that approach and access through alternative terms are not possible. The users looking for information through such alternative terms are not able to find the information though the information may be available in the file.

Assigned indexing is based on conceptual analysis of terms and words. The analysis is done to find out the concept and deciding the terms/words representing them and also the related concepts. It helps the user to reach to the required as well as related information. This assumes importance in view of the fact that the user may not be exactly sure of his/her information requirements. Even if he/she is sure of his/her requirements, he may not be able to express it exactly. Thus, a map of related concepts presented before him would help him to better understand, represent and reach to his required information.

2.3 INDEXING LANGUAGE

To understand the concept of indexing language it will be proper if we first dwell on the concept of language. Language is the vehicle for communication; it is a carrier for thought and plays an important role in communication. It enables the thought to flow from the source to the sink or from the origin to the destination. Communication used to take place even prior to the development of languages. Gestures were one of the ways by which it could take place, and these are still being used along with language for communication. Humans use a language that is different from what animals use to communicate. Humans also use different languages typical to their environmental and cultural factors.

From now on we shall use the term language to refer to only the language that is used by humans. The characteristics of a language are vocabulary and rules for their arrangement (syntax). The languages may be artificial and natural. Natural languages refer to our languages, which we normally use for communication, whereas, artificial are those that we have designed for a specific

purpose or are used in a specific sense or for limited use only. Shorthand is an example of this category which all of us have heard about. Similarly, we have examples of artificial languages in different disciplines e.g., in chemistry we have a language to indicate names of different elements, compounds and also the process of their transformation. Similarly, notation of a classification scheme is an artificial language.

2.3.1 Definition

Online Dictionary for Library and Information Science defines indexing as an artificial language consisting of subject headings or content descriptors selected to facilitate information retrieval by serving as access points in a catalogue or index, including any lead-in vocabulary and rules governing the form of entry, syntax, etc. In essence, it consists of a set of terms and devices for handling the relationship between the terms for providing index description. Sometimes it is also referred to as retrieval language. An indexing language is artificial in the sense that its semantics and syntax may be different from a natural language. It consists of a vocabulary and rules for admissible expressions.

2.3.2 Need and Purpose

To discuss about the need and purpose of an indexing language, it would be useful if we look into the need and purpose of a language. While using a language to satisfy the need for communication, there is no control in the language being used resulting in a lot of flexibility in conveying ideas. To understand this flexibility we have to see the process of flow of ideas in a language as illustrated in the Figure 2.1.

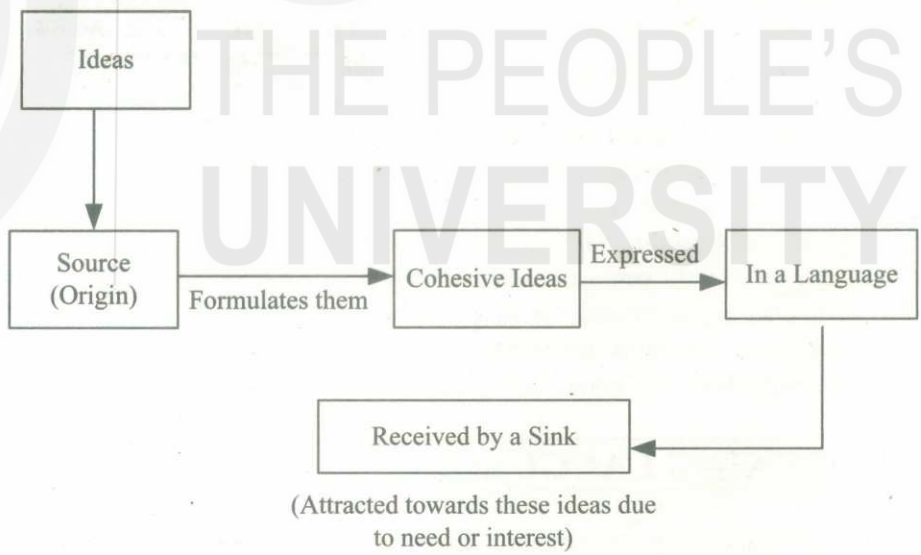


Fig. 2.1: Flow of ideas through languages

The ideas conveyed as shown above are not limited to a particular individual as a source and also not limited to a particular environment and location. The variations in environment and individuals result in flexibility in language. To express the same ideas, people use different languages, and also in different ways to communicate while using the same language. This varies even with the same set of people in different conditions and environment. This is termed as the flexibility of language, which has resulted in concepts like synonyms, homonyms, direct-indirect speech, active-passive voice, etc. Same ideas being expressed by different expressions or the same expression representing different ideas can be meaningful due to the context and interactive/two-way/real time/face to face communication. It may convey a wrong message otherwise. Therefore, in formal discussions, in

the vocabulary of a subject/discipline, terms are defined in advance and used in the same sense. Natural language is characterized by vocabulary and grammar augmented by the use of articles, prepositions, conjunctions, etc. In indexing, the concepts are indexed by essential words/ terms wherein the ancilliaries like prepositions, conjunctions are not used. Thus, for a compound/complex subjects, it is necessary that order of the terms (citation order) are prescribed. Further, natural language suffers from the problems of synonym, homonym, etc. While indexing, care should be taken to solve this problem so that all the documents on a particular topic are retrieved. The indexing system should, therefore, have provision not only to take care of the problems of ambiguity of terms and arrangement of terms, but should also show the relationships among different concepts so that complete retrieval effectiveness is achieved.

Thus, indexing language is necessary because:

- different authors in documents may express same concepts in different ways.
- different concepts may be expressed by same terms.
- users may be interested in these concepts. They need to be informed that their documents exist discussing the concepts specifically.
- users might be interested to access documents on related concepts also, thus, related concepts need to be brought together showing their relationships.

The purpose of an indexing language, therefore, is to:

- express the concepts discussed in documents;
- show the relationship among different concepts; and
- help in depicting a panoramic view of the related concepts.

Self Check Exercises

1) Define indexing. Enumerate its need.

2) Distinguish between assigned and derived indexing.

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

.....

.....

.....

.....

.....

.....

2.3.3 Characteristics

We have seen that the purpose of an indexing language is to express the concepts of documents in an artificial language so that users are able to get the required information. The indexing language does this by depicting the relationships among the different related concepts. Thus, an indexing language consists of elements that constitute its vocabulary, rules for admissible expressions (i.e. syntax) and semantics. An indexing language should, therefore, have:

- semantic structure
- syntactic structure
- syndetic structure.

Semantic Structure

Semantics refers to the aspects of meaning. In the context of an indexing language, two kinds of relationships between concepts – hierarchical and non-hierarchical can be identified. The hierarchical relationships may be Genus-Species and Whole-Part Relationships. The Non-hierarchical relationships may be Equivalence or Associative relationships.

Hierarchical Relationships: It is a permanent relationship.

a) Genus-Species (Example: Telephone is always a kind of Telecommunication)

b) Whole-Part (Example: Human Body - Respiratory system)

c) Instance (Example: Television - Phillips TV).

Non-Hierarchical Relationships: It may be of two kinds – Equivalence and Associate.

a) **Equivalence**

- Synonym (Example: Defects - Flaws)
- Homonym (Example: Fatigue (of metals), Fatigue (of humans);

b) **Associate**

It refers to the relationships in which concepts are semantically related but do not necessarily belong to same hierarchy (e.g. Weaving and cloth).

Syntactic Structure

As you know the word syntax refers to grammar. In the context of indexing language syntax governs the sequence of occurrence of terms in a subject heading viz., for the title export of iron, it may be Iron, Export or Export, Iron.

Syndetic Structure

To show the relationships described at semantic structure, syndetic structure should be built in indexing language (viz., *see, see also; use, use for*). Syndetic structure in the indexing language aims to link related concepts otherwise scattered and helps to collocate related concepts. It guides the indexer and the searcher to formulate index entries and to search for his/her information.

2.3.4 Types

The two types of indexing discussed earlier, i.e., derived and assigned, are different so far as the representation of contents of documents is concerned. The representation of concepts may or may not be the terms used by the author. Likewise the indexing languages may also be of different kinds, viz.: Natural indexing language, Free indexing language and Controlled indexing language. Natural language indexing uses the same vocabulary as those used by the author to represent the concepts. It is used in derived indexing. Free indexing language makes use of all possible terms to index documents irrespective of their use by authors. These would include all possible forms including synonyms, technical versus popular terms, words used in different areas, etc. Controlled language limits the use of terms based on the system used. It is used for assigned indexing. There are different examples of controlled languages, e.g., Lists of Subject Headings, Classification schemes, Thesauri, Thesaurofacet and Classaurus.

2.4 VOCABULARY CONTROL

Indexing may be thought of as a process of labelling items for future reference.

Considerable order can be introduced into the process by standardising the terms that are to be used as labels. This standardisation is known as vocabulary control, the systematic selection of preferred terms.

Lancaster [1986] suggests that the process of subject indexing involves two quite distinct intellectual steps: the 'conceptual analysis' of the documents and 'translation' of the conceptual analysis into a particular vocabulary. The second step in any information retrieval environment involves a 'controlled vocabulary', that is, a limited set of terms that must be used to represent the subject matter of documents. Similarly, the process of preparing the search strategy also involves two stages: conceptual analysis and translation into the language of the system. The first step involves an analysis of the request (submitted by the user) to determine what the user is really looking for, and the second step involves translation of the conceptual analysis to the vocabulary of the system. Thus there is a close resemblance between indexing and search process.

There are two major objectives of vocabulary control in an information retrieval environment:

a) to promote the consistent representation of subject matter by indexers and searchers, thereby avoiding the dispersion of related materials. This is achieved through the control (merging) of synonymous and near synonymous expressions and by distinguishing among homographs;

b) to facilitate the conduct of a comprehensive search on some topic by linking together terms whose meanings are related.

Lancaster [1986] further adds that indexing tends to be more consistent when the vocabulary used is controlled, because indexers are more likely to agree on the terms needed to describe a particular topic if they are selected from a pre-established list than when given a free hand to use any terms they wish. Similarly, from the searcher's point of view, it is easier to identify the terms appropriate to information needs if these terms must be selected from a definitive list. Thus, controlled vocabulary tends to match the language of indexers and searchers.

A large number of documents have appeared covering the details of various vocabulary control tools [for example, Aitchison and Gilchrist, 2000]. There are also standards such as the British Standards (BS 5723 and BS 6723), International Standards (such as ISO 2788 and ISO 5964), and UNISIST guidelines [1980, 1981].

A number of vocabulary control tools have been designed over the years: they differ in their structure and design features, but they all have the same purpose in an information retrieval environment. Availability of vocabulary control helps both the indexers, i.e., people who are engaged in creating document records, particularly those who create subject representation for the documents (by using keywords, in a post-coordinate system, for example), as well as the end-users in the formulation of their search expressions.

From the earlier discussion it should be clear that a natural language system suffers from varieties of problems in the context of development of an index file. Thus, the need for control of the vocabularies arises. A controlled vocabulary refers to an authority list of terms showing the inter-relationships and indicating the ways in which they may be combined to represent specific subject of a document. A certain degree of semantic structure is introduced in the controlled vocabulary so that terms whose meanings are related may be brought together or linked in some ways. This semantic structure is incorporated by means of (a) controlling the synonyms, word forms, etc. and distinguishing homographs for consistent representation of the subject of the documents; and (b) providing

mechanism to link the hierarchical and non-hierarchical terms that are related semantically to facilitate comprehensive search. Different techniques of vocabulary control have been adopted in the tools have List of Subject Headings (LSH), Thesaurus, Thesaurofacet, etc.

2.4.1 Controlled Vs. Natural Language Indexing

As Aitchison and Gilchrist [2000] pointed out, the differences between natural language indexing and controlled indexing are as follows:

Table 2.1: Comparison Between Controlled and Natural Language Indexing

Natural Language	Controlled Language
High specificity gives precision.; excels in retrieving individual terms – names of persons, organisations, etc.	Lacks specificity, even in detailed systems
Exhaustivity gives potential for high recall	Lacks exhaustivity; cost of indexing to the level of natural language is high; terms may be omitted due to indexer errors
Up-to-date; new terms are immediately available	Not possible to make it up-to-date; new terms are added to the thesaurus only at certain intervals
Words used by the authors are used in indexing; there is no provision for misinterpretation	Indexers may misinterpret some words; there is a risk of loss of terms and/or their proper meaning/use
Natural language used by the searcher is used for matching	The searcher may have to use artificial language (the controlled vocabulary)
Input cost is low since there is no human intervention	Input cost is high since human intervention may be required for translation of terms into appropriate controlled vocabulary terms
Language incompatibility can be overcome due to the use of natural language, and thus data exchange and cross-database searching becomes easier	The databases will have to use the same vocabulary control tool in order to be compatible
Intellectual effort is placed on searcher. Problems arise with terms having many synonyms and near-synonyms	Eases the burden of searching: – controls synonyms and near synonyms and leads to specific preferred terms to broaden search – qualifies homographs – provides scope notes – displays broader, narrower and related terms – expresses concepts elusive in free text
Syntax problems: there is a danger of false drops through incorrect term association	Overcomes syntax problems with compound terms and other devices
Exhaustivity may lead to loss of precision	Loss of precision can be avoided by selecting specific/narrower terms while searching

2.4.2 Methods of Vocabulary Control

Methods of vocabulary control can be discussed under the different heads that lead to various options/flexibility in languages. These are:

a) ***Semantics***

i) **Synonyms:** Synonyms lead to the same concept being denoted by different terms. For control, one of the terms is used/accepted, the rest of the terms are linked to the accepted term. The terms are connected through the

connecting device *See* or *Use* or *Used For (UF),* e.g., child medicine and pediatrics are synonyms. If child medicine were the accepted term, then there would be a link.

Pediatrics
***See* Child Medicine or**
Use Child Medicine
and another link
Child Medicine
UF Pediatrics

ii) **Variant Spellings:** A term may have variant spellings. One of these is chosen, and the other spelling would be linked to it, e.g.

Catalog
***See* Catalogue or**
Use Catalogue
And another link
Catalogue
UF Catalog

iii) **Homonyms:** Homonyms are concepts where the same term may have different meaning and cause a problem in understanding the concept behind a term. Vocabulary control is achieved by providing the context in brackets along with the term, e.g.,

Bridge (Road)
Bridge (Game)

b) ***Syntactics***

All the points, discussed above under headings (i) to (iii), are related to the semantics of the language. Thus, they help to control the variations in semantics. Similarly, the flexibility in syntax has also to be controlled. These occur due to headings occurring in compound terms and phrases.

i) **Compound Terms:** Terms representing a subject may occur as compound terms in a combination of a noun and an adjective. These may be expressed in different ways e.g. Academic libraries could also be expressed as Libraries in academic institutions or as Libraries, academic.

If we follow the first option, it becomes a direct heading, otherwise an inverted heading. Both have their own advantages. Direct heading provides the traditional use option whereas the inverted heading helps to collocate all material/literature on the subject at one place. But one of these has to be followed, linking to the other by *see* or the *use* reference for the help of the user.

ii) **Phrase Headings:** There may be subjects that need to be expressed in phrases using conjunctions or prepositions. There are different ways in which these can be expressed, e.g., Role of libraries in societal development. It could also be expressed as Libraries- Role in societal development, or Societal development, role of libraries in.

Self Check Exercise

3) Discuss the issues of vocabulary control.

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

.....

.....

.....

.....

.....

2.5 CLASSIFICATION SCHEMES

Though classifications schemes have been primarily developed to mechanize arrangement of shelves, it is one of the examples of indexing languages. It makes use of artificial numbers instead of language to represent concepts. These enable the indexer to show the hierarchical relations as well as to put them in order. The equivalence and associative relationships are also shown by various mechanisms in the schemes. The citation order of classification schemes reflects the characteristics of syntax of an indexing language. Classificatory principles are always involved in indexing whatever method we may adopt. Indexing involves two processes, analysis of subjects of documents and their representation, which is true to classification also. The only difference lies in the method of representation.

Classification schemes belong to different categories, viz. enumerative, analytico synthetic and faceted. The degrees of enumerativeness and facetedness vary from partial to full. Library of Congress scheme belongs to the former category and Colon Classification to the latter. Dewey Decimal Classification (DDC) began as an enumerative scheme but now it has incorporated quite a good degree of facetedness. Subjects of different kinds varying from simple to complex are all enumerated along with their notation in an enumerative scheme. In a faceted scheme only the possible facets are enumerated, a subject can only be represented by joining the facets after analysis of the subject. The Universal Decimal Classification scheme is an example of analytico synthetic scheme which is not freely faceted. You have already learnt the DDC and CC in detail in your BLIS Programme. The features of the these schemes particularly Universal Decimal Classification Scheme have been discussed in the next Unit.

2.6 SUBJECT HEADINGS LISTS

Subject headings are provided in the catalogue entries to provide subject access to information. Cataloguers have at their disposal different Lists of Subject Headings from which they can assign subject headings to the documents that they catalogue. Library of Congress List of Subject Headings (LCSH) and its abridged form, the Sears List of Subject Headings (SLSH) are the two main lists used in libraries. The Subject Headings Lists show the semantic aspects of indexing language through *See* and *See Also* references. The relationships - Hierarchical and Associative are accommodated under *See Also* reference. The syntax is handled by various instructions in the lists. Though subject headings lists were developed from card catalogue and pre-coordinate indexing systems (in particular), due to development of computerised information retrieval system, these are gradually transforming the thesaural structure, which is developed for post coordinate systems. You have learnt in detail the SLSH in your BLIS programme. The features of LCSH and SLSH are discussed in this Unit in subsequent section.

2.7 THESAURUS

Thesaurus is another example of an indexing language providing vocabulary control to be used to index and search information. Thesaurus was designed to function for post-coordinate systems, also helping the searcher to conduct the search in systematic way.

The oldest living example of a thesaurus is the Roget's Thesaurus given by Peter Mark Roget in 1852. It was developed to provide alternate terms for a given concept and is divided into two parts viz., classified and alphabetical. The classified part has certain categories further subdivided into subdivisions under which are placed the different words. The words are assigned to different grammatical forms like noun, verb, etc. The other part is the alphabetical part consisting of the words arranged alphabetically with the reference to their category numbers linking them to the classified part. Though it is named as Thesaurus, it is completely different from an information retrieval thesaurus, as we use the term today, both in purpose, functions and structure. IR thesaurus basically serve as a controlled vocabulary used for indexing and searching of information in an ISAR system.

2.7.1 Definition

Online Dictionary for Library and Information Science [2005] defines a thesaurus as an alphabetically arranged lexicon of terms comprising the specialized vocabulary of an academic discipline or field of study, showing the logical and semantic relations among terms, particularly a list of subject headings or descriptors used as preferred terms in indexing the literature of the field. Brownson first used this term in the context of information retrieval.

UNISIST defines it in terms of its structure and function: "In terms of function, a thesaurus is a terminological control device used in translating from the natural language of documents of indexers or users into a more constrained "system language" (documentation language, information language). In terms of structure, a thesaurus is a controlled and dynamic vocabulary of semantically and generically related terms which covers a specific domain of knowledge."

2.7.2 Functions

The functions of a thesaurus are:

- to provide a panoramic view of a subject/field showing relations among its constituents to help the indexer to assign descriptors to documents and the searcher to access them;
- to provide a standard vocabulary to the indexers for a subject/field;
- to show the relationships existing among concepts that could help searchers to narrow down or broaden their searches for effective retrieval;
- to provide a map of concepts in a subject/field to enable the indexer/ searcher identify different concepts which he could not have known otherwise, in many cases.

The difference between a list of subject headings and thesaurus lies in the function and structure. Subject headings lists were developed to suit the subject catalogue and also to use in the pre-coordinate indexing systems. Whereas Thesaurus has been developed in the context of post-coordinate indexing systems.

Thesauri have been designed in specific fields/areas by international organisations other than libraries. Some examples include: Thesaurus of Engineering and

Scientific Terms (TEST), ERIC Thesaurus for Education, Thesaurus of American Psychological Association (APA), Unesco Thesaurus, etc.

MICROFORMS	
SN	A miniature of replica of document
UF	Micro copies
BT	Data media
NT	Micro transparencies Micro-opaques
RT	Microphotography

Thesaural Display

2.7.3 Construction

Thesaurus construction is a very specialized activity. Anyone involved in its construction should have a sound knowledge of the subject and should be logical and have organisational capabilities. The steps for construction of a thesaurus are as follows [Lancaster, 1985]:

a) **Need Analysis**

While designing the thesaurus need analysis should be done first, whether it is really needed or not. There may be existing thesaurus on similar subjects. It is necessary to see whether it may meet the need. In some cases, an existing thesaurus can be modified to suit the needs. If it is felt that a thesaurus needs to be constructed then following steps to be followed.

b) **Gathering of Terms**

The terms to be included are to be collected first. Two approaches can be followed in this process. In the top-down approach (deductive approach), a committee identifies the terms and subdivide them from the top to down. The problems, which may be faced are that it is difficult to think of all categories or hierarchies of a concept and the characteristics used to divide the genus may not suit the users needs.

In the empirical (bottom-up) approach, terms are correlated from various sources and a category or hierarchy is formed only if it appears to be useful. The terms are collected using two principles - Principle of Literary warrant and Principle of User Warrant. In the former the logic is that a term justifies its inclusion if it is used in literature of the subject. The method is to go through abstracting sources, reference sources, periodical articles, etc. In the later case, users/ subject specialists may be consulted to gather the terms. However, the combination of the two yields better result.

c) **Organisation of Terms**

Once the terms are collected, these are to be organised into major categories and into hierarchies within the categories. Useful inter-hierarchical relationships should also be delineated.

d) **Organisation into Hierarchies**

Once the categories are identified, the next stage is to organize each term into hierarchies.

e) **Creation of Alphabetical Thesaurus**

Once the hierarchies are established, the classification is inserted to create alphabetical thesaurus. Each term becomes an entry and its hierarchical relationships are denoted by BT and NT. All the BT and NT terms should reciprocate. Similarly the non-hierarchical relationships are shown through *use, used for* and *related terms (RTs)*. Normally, one step up and one step down is followed.

f) **Presentation of Thesaurus**

Each block of entries are arranged according to requirement. It may be alphabetical, systematic (to complement) or graphic.

g) **Evaluation**

Once the thesaurus is compiled it needs to be evaluated to assess its retrieval effectiveness. This may be done according to norms and procedures as discussed in Unit 5 of the Course.

h) **Maintenance**

Once a thesaurus is developed, it should be maintained properly. New terms need to be added or deleted as the case may be. This has to be done continuously.

i) **Use of Computers**

The collection of terms as mentioned earlier is very tedious and time consuming. Computers can be effectively used in gathering of terms. Terms can be derived from machine readable databases through the use of statistical techniques.

Construction of thesaurus is largely an intellectual activity as far as delineating the relationships of terms is concerned. Once the terms are organised into facets and hierarchies, the use of computers can be useful. The computers can print/ display. Further, computer readable thesaurus data can be used for photocomposition to produce the print version. The most important application of computer is in the maintenance of thesaurus. The addition and deletion of terms may be done very effectively through the use of computers. Many thesaurus are now available in computer readable form and linked with the databases. While searching, the system automatically converts the terms into the terms of thesaurus and conducts the search.

Self Check Exercise

4) Enumerate the structural differences between a Thesaurus and a List of Subject Headings.

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

.....

.....

.....

.....

.....

2.8 THESAUROFACET

You must have noticed that an alphabetical thesaurus provides block of terms of a concept in which the relationships of terms with the concept are shown. It cannot show the complete hierarchy of subjects at one place. To overcome such problem a different type of tool has been developed, known as Thesaurofacet. Thesaurofacet has been developed by Jean Aitchison of English Electric Company. It is basically an integration of a faceted scheme of classification and thesaurus. The first part is a faceted scheme and the second part is an alphabetical thesaurus. The terms in the system appear once in schedule and once in the thesaurus. The link between the two is the notation or class number.

2.9 CLASSAURUS

Classaurus is another vocabulary control device devised by G. Bhattacharyya for POPSI indexing system. It is a faceted systematic scheme for hierarchical classification incorporating all the features of a thesaurus. It does not have two separate sections viz., classification schedule and thesaurus. It has only one main section consisting of separate schedules for different facets, but the schedules incorporate within themselves the features of a classification scheme and thesaurus. Because of this it has been called Classaurus. The fundamentals of classaurus forms part of a General Theory of Subject Indexing Language. The index of a classaurus is an alphabetical index.

2.10 SEARS LIST OF SUBJECT HEADINGS (SLSH)

This List owes its name to its originator Minnie Earl Sears who gave its first edition in 1923 as List of Subject Headings for Small Libraries that was based on the list of subject headings used by nine small well-catalogued libraries. She edited the List till 1933 when it was in its 3rd edition. The name of the list was changed to the present one since its 6th edition. Small Libraries was removed from the title as medium sized libraries also started using it. Sears was added to the title in recognition to the contribution of Minnie Earl Sears. It is in its 18th edition, which was published in 2004. The feature of this edition is Principles of the Sears List, which enlists the theoretical principles of subject headings used in the List as well as general principles of subject cataloguing. The principles guide the indexer in formulating new headings or subdivision of headings not provided for in the List. The List follows the LCSH, however there is a difference as stated in the introduction: "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."

SLSH has got a new face since the 15th edition, which was published in 1994. Since then editions have been coming quite regularly viz., 16th in 1997, 17th in 2000, and 18th in 2004.The new face is due to the change in format that follows the NISO standards for thesauri. Earlier references, *See*, *See also*, *x*, and *xx* have been replaced with USE, NT, BT, RT and UF. Headings that have been replaced have been appended with the phrase [Former heading]. Another significant change came in the 16th edition when new headings were added for other religions to reduce the Christian bias. Some additions were done in the headings for ethnographic divisions, computers, personal relations, politics and popular culture in the latter editions.

Examples Illustrating Use of SLSH

SUBJECT	SUBJECT HEADING
1) Tourist places in India	India – Description and travel
2) Marriages in Japan: A study of customs	Marriage customs and rites – Japan
3) Pulse polio campaign in India: Public health awareness	Vaccination – India
4) Directories of libraries in Delhi	Libraries-Delhi-Directories
5) Surgery of the human heart	Heart-Surgery

Headings in LCSH Vs. SLSH

LCSH	SLSH
1) Library education - India	Library education - India
2) Library orientation - India	Bibliographic Instruction
3) Blood - Cancer - Treatment	Blood Cancer - Treatment
4) English language- Dictionaries	English language- Dictionaries
5) India - Foreign relations - Pakistan	India-Foreign relations - Pakistan

Self Check Exercise

5) Assign subject headings to the following subjects using LCSH and SLSH

a) Cataloguing of manuscripts in art libraries

b) Homeopathic treatment of skin diseases

c) Grammar of Sanskrit language

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

.....

.....

.....

.....

.....

2.11 LIBRARY OF CONGRESS LIST OF SUBJECT HEADINGS

Library of Congress has been providing subject headings in its catalogue since 1898. The libraries using L.C. cards requested it to publish these headings for other libraries to use. L.C. started publishing these, when its first edition was published between 1909 and 1914. SLSH is based on Library of Congress List of Subject Headings (LCSH) designed for small and medium sized libraries.

LCSH was published for the first time as "Subject Headings used in the Dictionary Catalogues of the LC" between 1909 and 1914. Later on supplements were published followed by the second edition issued in 1919. The list is in its 26th edition at present, which was published in 2003. It is in five volumes. The present editor of the list is Ronald A. Gowdreas. The list is generated from a database accumulated since its inception. The idea of the size of the list can be had from the fact that, it has 2.7 lakh records compared to 2.63 lakh records in LCSH 25.

There are two types of headings in LCSH, those in bold face type and those in normal type. The bold face type headings are accepted to be used as subject

headings by catalogues, others cannot be used as headings, and they guide the users to formulate the appropriate subject headings, e.g.,

Fighting

Use **War**

Fighting is an example of non-bold face type heading which is not an acceptable subject heading. This entry guides the user to choose **war** as the heading.

Components of an entry

An entry in LCSH may consist of the following elements:

Scope Note (SN): It provides the meaning or context of a heading. The scope note also gives an indication of the area of application of the heading, e.g.,

Art

(Here are entered general works on the visual arts. Work on the arts in general, including the visual arts, literature, and the performing arts, is entered under Arts).

Thus, the scope note sets the context and distinguishes the heading Art from Arts. Scope note is not required in all entries. It is only required when there exists an overlap in entries.

Directing Elements

An entry may have the following directing elements:

USE, UF (Used For), BT (Broader Term), NT (Narrower Term), RT (Related Term) and SA (See Also).

Directing elements are used to interconnect related subject headings. Related subject headings are interconnected to help the indexer and user/searcher. The indexer might have thought of a subject heading to assign to a document that doesn't exactly represent the subject of the document. To help him to refine his subject heading, the directing elements are used. Similarly for the user, he may use a subject heading to find out the information required by him, which is not the appropriate heading. The directing elements guide him in refining his heading to reach his required information.

A subject is related to another subject in any of the following three relational ways:

Equivalence

Hierarchical

Associative

a) ***Equivalence Relations***

These relations exist between subjects due to word forms, e.g., synonyms, technical versus popular terminology, abbreviations or acronyms versus their full forms, variations in spellings, multiword terms, etc.

Since one among the variations of a term is to be used as an acceptable heading, it is connected to others through USE reference, e.g.,

Hindu Gods

USE **Gods, Hindu (May sub Geog)**

Adult offspring

USE **Adult children**

Carcinoma
USE **Cancer**
Iconographs
USE **Art**
Injuries
USE **Accidents**
Midday
USE **Noon**
Cataloging
USE **Cataloguing**

b) ***Hierarchical Relationships***

Subjects related to each other as whole part of whole type are designated as hierarchical relationships. The directing elements used to connect. Those are BT and NT, their full forms being broader term and narrower term respectively.

Examples :

Gods, Hindu (*May subd Geog*)
BT Hinduism
Adult Children (*May subd Geog*)
BT Adulthood
Children
NT Abusive adult children
Adult children living with parents
War
BT International relations
Cancer
BT Tumors

c) ***Associative Relations***

Concepts not related to each other as equivalent or hierarchical are said to have associative/attentive relation. Lancaster categorises these relations as:

Coordination
Genetic
Concurrent
Cause and Effect
Instruments
Materials
Similarity

Two concepts are said to be coordinate when they are species of the same level of a genetic concept, e.g.,

Heating
RT Ventilation
Hearing
RT Deafness
Economic Policy
RT Social Policy

Genetic relationships

Genetic relationships may not be hierarchical coordinate e.g.

Sons
RT Daughters

Concurrent relations arise out of two or more activities occurring at the same time/concurrently.

Cause and Effect relations involves terms that are related to each other as cause and effect of each other e.g.,

Teaching
RT Learning
Education
RT Learning and relationship
Alcoholism
RT Alcoholics
Advertising
RT Publicity

Instruments: Such relationship involves concepts that devote an action and the instrument used for doing it. e.g.,

Teaching
RT Projector
Temperature
RT Thermometers
Public opinion
RT Press
Public health
RT Sanitation

Materials:
Ceramic Industry
RT Ceramics
Cells
RT Protoplasm
Money
RT Gold
Silver

Similarity:
Singing
RT Voice
Single Women
RT Divorced Women
Religious Holidays
RT Fasting

General References

The references that we have discussed till now, are all specific references, i.e., they guide a user from one heading to another possible heading. There is another group of headings that are generic in nature and guide the user from one heading to a group of headings. The reference provides, as guidance, one or two possible headings under the generic reference. In such cases a reference may also be given to a subdivision the directing elements, i.e. SA, for examples:

War

SA warfare under ethnic groups; also subdivision wars under ethnic groups; and individual wars and battles.

This reference helps us to formulate a heading, e.g.,
India - History - Mysore War, 1790-1792

Adult education (may subd Geog)
SA subdivn Adult education

under individual Christian denunciation, e.g.,

Catholic Church - Adult education

Art (may subd Geog)

SA subdivn Art under names individual person who lived before 1400 under names of duties or legendary figures, and under headings of the type (topics) - (Subdivn), e.g.,

Mary Blessed Virgin, Saints - Art

Therefore, on the analogy, we may assign a heading Kabirdas, Saint - Art
SA references have also been given for form headings, which have been discussed under form subdivision.

Subdivisions

There is a provision of subdivisions in headings to enable a user to formulate specific headings. This subdivision may be of different levels that are indicated by the dashes prefixed to them. A single dash prefixed to a heading indicates that it is first level and two dashes prefixed indicate that it is a second level heading. However, while assigning headings, the indexer uses one dash only, e.g.,
India - History - Emergency Period

The subdivisions are of four types, viz., topical, form, chronological, and geographical.

Topical subdivisions are used to limit the heading to a particular aspect or in a particular context only. Form subdivision is used to specify the form in which a document is presented e.g., history, report, encyclopaedia etc. Such subdivisions are not added to headings in LCSH, their references are given in the list for adding such subdivision, e.g., Encyclopedias and dictionaries.
SA subdivisions Dictionaries or Encyclopaedias under subjects.

Chronological subdivisions are used to limit a heading by the period of coverage. Such Subdivisions are of two types, viz., a) Established headings under names of different jurisdiction, e.g., Philosophy, French - 18th century
Art, Chinese - TO 221 B.C.

Headings provided for United States with subdivision History, Economic conditions and Politics and government respectively. Analogous headings may be provided for other countries.

Geographical Subdivisions

Subject of some documents is limited in respect to the geographical area discussed e.g. Library Science in India. It is to be seen whether we can specify the geographical aspect, India, in the subject heading or not. We can specify the jurisdiction in those subject headings only if it is specified so. The indication for this is, (May Subd Geog). Any heading suffixed with such a phrase allows the indexer to subdivide the heading by a geographical jurisdiction. Those headings that do not have such a phrase attached to them should not be subdivided geographically, e.g., war, War (May Subd. Geog.) permits us to assign a subject heading:

War-India

However, **Art** - (May Subd. Geog.) allows us to assign. a subject heading **Art - India**

The introduction to LCSH states that those headings, which do not allow subdividing geographically, are under review (whether the geographical subdivision should be allowed or not). Thus, they should not be subdivided geographically, Geographical Subdivisions – state, country, province, district, should be indirect. Indirect subdivision means that the subject heading should be sub divisional by country and then the smaller area, e.g., for a subject 'Art in U.P.' should have a subject heading, Art-India-U.P. But there is deviation in the rule for United States, Great Britain and Canada. Thus, the subject heading for the subject 'Art in Detroit' would be 'Art-Detroit'. The reason for this difference is that, literature on subjects with subdivisions lower than country is collected by the names of countries, while it is done by names of states, districts, provinces etc. in case of US, Great Britain and Canada.

There might be headings which may be possible with more than one type of subdivisions. In such a case, form division comes at the end. Location of geographical subdivision depends on which topics can be subdivided by place, e.g., (from LCSH):

Construction industry (may subd Geog)
Finance
Law and Legislation (may subd Geog)
Government Policy (may subd Geog)
Mathematical models

In the context of India, these would result in the following possible subject headings respectively:

Construction Industry - India
Construction Industry - India - Finance
Construction Industry - Finance - Law and Legislation
Construction Industry - Govt. Policy - India

Form of Headings

Headings in LCSH adhere to the following:

a) Preference is given to one word headings over multiword headings;

b) Concepts are stated in singular and objects in plural.

c) Topic multiworded headings follow normal used order except in case of "headings with language, nationality or ethnic adjectives, headings qualified with three period, qualified by artistic style, headings with the adjective. Fossil and certain music headings."

Examples of inverted headings:

Gods, Hindu (may subd geog)

Art, Indic (may subd geog)

Infertility Male

d) Phrase headings are used in which conjunctions or preposition are present. They are present in normal word order generally. There are exceptions at some places where they have been inverted.

Examples:

Art and Science

Art and Society (may subd. Geog)

Science and Civilization

But **plants, Effect of the moon on.**

Examples Illustrating use of LCSH:

1) *Subject :* **Education for LIS in India**

Study and Teaching is a Floating subdivision and can be used as a subdivision under subjects and can be used as a subdivision under subjects for works on methods of study and teaching of those subjects. But there is an entry,

Library Science - Study and teaching.

USE **Library education** (*May subd. Geog*)

Therefore for India, heading will be

Library education - *India*

2) *Subject :* **User education in libraries in India**

Under user education there is a USE reference

USE **Library orientation**

The entry is Library orientation (*May sub. Geog*)

Therefore in the context of India, the heading will be **Library Orientiation- India**

3) *Subject :* **Treatment of blood cancer**

The entry under cancer is SA subdivision cancer under individual organs and regions of the body, e.g., **Foot - cancer.**

Therefore, the heading can be Blood-cancer.

Treatment of above, a floating subdivision – therefore the heading will be:

Blood - Cancer - Treatment

4) *Subject :* **Dictionary of English language**

Encyclopaedias and Dictionaries is a floating subdivision. The instruction under the heading is "Words of a specific language usually with definitions are entered under the heading of the language with subdivision Dictionaries."

Therefore, the heading for the subject will be: **English language - Dictionaries**

5) *Subject :* **Foreign relations between India and Pakistan**

For subject headings of a country, entries under United States can be used. Under United States, there is an entry,

United States - Foreign relations - Japan

On the same analogy, the heading will be

India - Foreign relations - Pakistan

2.12 SUMMARY

In this Unit an attempt has been made to introduce you to the need and concept of an indexing language. The types of indexing have been discussed briefly to justify the need and problems of indexing language. There is a discussion of the concept of vocabulary control followed by the types of vocabulary control devices. LCSH has been discussed at length. SLSH has also been described briefly (as it has been introduced to you earlier at BLIS level). Examples from both the lists have been discussed to clarify the concepts.

Thesaurus is form of vocabulary control devices. It has been introduced in the Unit with a discussion of its structure and organisation. It is followed by a explanation of the construction of thesaurus.

2.13 ANSWERS TO SELF CHECK EXERCISES

1) Indexing is the process of representing the subject of the documents to facilitate access and retrieval later. Indexing is needed because:
   - Different authors in documents may express same concepts in different ways.
   - Users may be interested in these concepts. They need to be informed that there exist documents discussing these concepts.
   - Users might be interested to access documents on related concepts also, thus, related concepts need to be brought together depicting their relationships.

2) Derived indexing refers to the method of indexing wherein only the terms used by the authors are used for indexing the documents. In assigned indexing, the terms used for indexing are based on the analysis of concepts behind the terms. The terms are analysed for relations existing between them.

3) Vocabulary control addresse to the following issues:
   - Semantics related to the roots of words and their meanings. These occur as: Synonyms, homonyms, variant spellings, popular versus technical terminology, abbreviations, acroynyms, etc.
   - Syntax is related to representing ideas by joining concepts. These occur as phrases and compound terms. In vocabulary control the variations in representing the same ideas are controlled. The alternative options are connected by references.

4) The structural difference between a LSH and a thesaurus is that a LSH is in one part whereas a thesaurus may also exist in two parts viz., classified and alphabetical. In a thesaurus, uniworded terms are used as these are meant for post-coordinate indexing. Relationships in the terms are depicted more exhaustively.

5) a) Cataloguing of manuscripts in art libraries - SLSH- Art Libraries; Cataloguing; Manuscripts

   LCSH- Art Libraries; Cataloguing of manuscripts

   b) Homeopathic treatment of skin diseases-SLSH - Homeopathy; Skin-Diseases-Treatment

   LCSH- Skin- Diseases - Homeopathic Treatment

   c) Grammar of Sanskrit language - SLSH and LCSH

   Sanskrit language - Grammar

2.14 KEYWORDS

Assigned indexing : The system of indexing in which terms are assigned by the indexer on the basis of the conceptual relations existing between them.

Derived indexing : The system of indexing in which terms are derived from the terms used in the text.

Thesaurofacet : Thesaurus which has a classified and alphabetical part.

Indexing language : A language used to describe and search subjects in an index.

Vocabulary control : The technique of using controlled language to describe terms in an information storage and retrieval system.

2.15 REFERENCES AND FURTHER READING

Aitchison, J. and Gilchrist, A. (2000). *Thesaurus construction and use: a practical manual*. 4th ed. London: Aslib.

Artherton, Pauline. (1977). *UNESCO handbook of information systems and services*. Paris: UNESCO.

BS 5723: 1987. *Guidelines for establishment the search of monolingual thesaurus*. London: British Standard Institution.

BS 6723: 1985: ibid.

Chowdhury, G.G. (2000). *Modern information retrieval*. London: Clive Bingley.

Foskett, A.C. (1996). *The subject approach to information*. 5th ed. London: Library Association Publishing.

Ghosh, S.B. and Satpathi, J.N. (eds.). (1998). *Subject indexing systems: concepts, methods and techniques*. Calcutta: IASLIC.

Guha, B. (1983). *Documentation and information*. Calcutta: World Press.

ISO 2788: 1986. *Guidelines for establishment and development monolingual thesaurus*. Geneva: International Organization for Standardization.

ISO 5964:1985. ibid.

Lancaster, F.W. (1985). *Thesaurus construction and use: a condensed trend*. Paris: UNESCO (PG1-85/WS 11).

Lancaster, F.W. (2003). *Indexing and abstracting in theory and practice*. 3rd ed. London: Facet Publishing.

Reitz, Joan M. *Online dictionary for library and information science*. <http://lu.com/odlis/odlis_i.cfm> (Accessed on 3rd, May 2005)

Sparck Jones, Karen and Willett, Peter. (eds). (1997). *Readings in information retrieval*. San Francisco, California: Morgan Kaufmann Publishers.

UNISIST (1981). *Guidelines for establishment in development of monolingual thesaurus*. 2nd ed. Paris : UNESCO.

ignou THE PEOPLE'S UNIVERSITY

UNIT 3 INDEXING LANGUAGES – PART II: CLASSIFICATION SCHEMES

Structure

3.0 Objectives

3.1 Introduction

3.2 Dewey Decimal Classification (DDC) Scheme

- 3.2.1 New Edition – DDC-22
- 3.2.2 Changes in the DDC-22
- 3.2.3 WebDewey
- 3.2.4 Popularity and Problems

3.3 Universal Decimal Classification (UDC) Scheme

- 3.3.1 Historical Background
- 3.3.2 The UDC Consortium
- 3.3.3 English Editions of UDC
- 3.3.4 Features and Structure of the UDC
- 3.3.5 Filing Order
- 3.3.6 Citation Order
- 3.3.7 Qualities of the UDC Scheme

3.4 Library of Congress Classification (LCC) Scheme

- 3.4.1 The System
- 3.4.2 The Use, Revision and Future

3.5 Colon Classification (CC) Scheme

- 3.5.1 Features of 7th Edition of CC
- 3.5.2 Notation of CC-7
- 3.5.3 Revision, Use and Status

3.6 Bibliographic Classification (BC) Scheme

- 3.6.1 The Second Edition, the BC-2
- 3.6.2 Special Features of BC
- 3.6.3 Structure, Auxiliary Tables and Use

3.7 Library Bibliographical Classification (BBK/LBC) Scheme

- 3.7.1 Structure of the BBK

3.8 Broad System of Ordering (BSO) Scheme

- 3.8.1 Structure
- 3.8.2 Use

3.9 Special Classification Systems

- 3.9.1 Benefits
- 3.9.2 Scope
- 3.9.3 Approaches
- 3.9.4 Some Examples

3.10 Trends of Research in Classification Systems

3.11 Summary

3.12 Answers to Self Check Exercises

3.13 Keywords

3.14 References and Further Reading

ignou THE PEOPLE'S UNIVERSITY

3.0 OBJECTIVES

In the previous Unit, you have learnt about indexing languages and their characteristics where you have also come to know about the classification schemes as an example of indexing language. In this Unit we will discuss a few classification schemes particularly Universal Decimal Classification Scheme which are available for organisation of knowledge in libraries.

After reading this Unit, you will be able to:

- know the major existing general classification systems being used today;
- understand their structure, main classes, notation and characteristic features, limitations and problems;
- understand their historical contributions to the science of library classification;
- be in a position to choose a suitable classification system for your library;
- understand the special classification systems for micro-subjects or specialised forms of information; and
- have a peep into the state-of-the-art in the field, and foresee future trends.

3.1 INTRODUCTION

You have already studied the importance of classification system in intellectual organisation of information (IOI) in Unit 1. In the following sections, we shall discuss features of a few general classification systems to give an idea how the various schemes try to organise the universe of knowledge. You know that in 1876 the era of modern classification started with the publication in USA of what is now universally known as Dewey Decimal Classification (DDC). This scheme of classification provided a new and efficient system of arranging books on the library shelves, and solved many of vexed questions then faced by the librarians. It started a new era of library classification, nay, of modern librarianship in general. On this account, the inventor of the DDC, the famous US librarian Melvil Dewey (1851-1931), is deservedly called the Father of Modern Librarianship. In the wake of the DDC many other classification systems have come up either to improve upon the Dewey's shortcomings, or to try alternative approaches to organise knowledge or to apply another set of notational system. Since then, a number of general classification systems have emerged. The trend continues, though at the moment no general classification system seems in the making. Some of the general classification systems, e.g. "Expansive Classification" (1893) designed by another famous US librarian Charles A Cutter (1837-1903) and the "Subject Classification" (1906) by a British librarian James Duff Brown (1862-1914) are now dead systems because of the inherent faults or neglect, or of both. International Classification (1961) by an US librarian Fremont A Rider (1885-1962) remains ignored in the library world. Though designed comparatively recently, it is a primitive system in method and details. There are some other general classification systems made and used locally. We will study here the major living general classification systems in the subsequent sections.

The special classification system is another important tool for intellectual organisation of information in some special libraries and information centres as well as in some information systems, that deal with micro-subjects or specialised forms of documents. Some individuals and organisations have developed a number of special classification schemes. These schemes are being used in some information centres and information systems to cater efficient information management. This Unit illustratively describes features of general and special classification systems.

3.2 DEWEY DECIMAL CLASSIFICATION (DDC) SCHEME

The first and the most popular general classification system now was first published in 1876 by Melvil Dewey (1851-1931). The system divides the entire universe of knowledge into ten main classes denoted by pure and simple notation of Indo-Arabic numerals used decimally. Decimal notation helps to incorporate new subjects indefinitely at any point in the hierarchy. Position of documents is relative, not fixed. Its other important contributions are : its Relative Index, and division of knowledge by 'Discipline'. It is available in three official versions of varying details namely Full edition DDC-22 (2003), Abridged DDC-14 (2004) and the electronic version called WebDewey. You have already studied DDC in your BLIS programme. Here we will be discussing only the changes which have taken place since DDC-19.

3.2.1 New Edition — DDC-22

You have already studied the DDC 19th edition in BLIS programme. The DDC is now in its 22nd edition published in September 2003.

> Dewey, Melvil: *The Dewey Decimal Classification and Relative Index*/ devised by Melvil Dewey. 22nd edited by Joan S. Mitchell, et al. Albany, N.Y.: Forest Press/OCLC, 2003. 4v. ISBN 0-910608-70-9.

In 1988 the Forest Press was acquired by the world famous OCLC at Dublin, Ohio, USA. Earlier, the Forest Press was a subsidiary of the Lake Placid Club Educational Foundation established in 1896 by Melvil Dewey himself. DDC-22 comprises four volumes.

Vol-1 Introduction Tables

Vol-2-3 Schedules

Vol.4 Relative Index. Manual on the Use of DDC

It has been computer edited by Editorial Support System (ESS) installed in 1984 by the DDC editorial office at the Library of Congress, Washington DC (USA). The 20th ed (1989) was the first edition generated from an online database. The ability to typeset, design, edit, modify and develop classification systems through computers has many advantages. A major advantage of the ESS is that the system can be updated minute by minute; and the production time has been reduced to six months from two years. It has enormous implications, namely on the frequency of new editions, annual electronic version, cost, and format.

3.2.2 Changes in the DDC-22

DDC-22 (2003) has been bit conservative in revision as no drastic change has been made, or earlier so called phoenix schedules have been added. Nevertheless, it has many updates, with many changes here and there. Expansions of many old topics are there along with a few structural changes. The major updates are:

004-006 Computer Science; 200 Religion further reduces the Christian bias. Now specific aspects of (comparative) religion are shifted to 201-209 from 291. Numbers for primitive religions of Blacks have been expanded. 305-306 Social Groups have been restructured. 340 law is now improved. 518 lying vacant since 1950 has been filled with numerical analysis. 543.1-.8 Analytical chemistry aspects, lying scattered earlier, have been brought together here. Class 610 gets new caption "Medicine and Health". It has also been updated using current terminology. Historical periods have been updated in 930-990 class. Amongst the Tables (vol. 1), Tables 1 and 3 have been updated. Tables 5 and 6 have expanded provisions

for American natives (Indians) and languages. Table 7 has been scrapped and merged in the schedules. This will not alter any class number belonging to persons. It is claimed by the editors that DDC-22 enhances the accuracy of classification better than previous editions.

3.2.3 WebDewey

Dewey for Windows (DfW) CD-ROM version of DDC-21 is no more available for DDC-22. Instead the electronic version called WebDewey can now be accessed only on the Internet (of course through subscriber's licence). Being web based enhanced version of the DDC-22 database, it contains many more innovative features with many access points and display options. It is web based enhanced version of the DDC database with online delivery that is updated continuously. It is compatible with Microsoft Internet explorer, version 4 or higher, and Netscape navigator version 4.x in 300 MHz Pentium II with 64 MB RAM. It is part of the OCLC suite of cataloguing and metadata services that the OCLC offers through OCLC connexion services (http://connexion.oclc.org).

3.2.4 Popularity and Problems

As far as its defects are concerned, the DDC, as ever is saddled with the problem of simple and primitive notation unable to provide placement to new subjects at proper places; lengthy class numbers; and the Western bias. Despite its defects and heavy criticism, it continues to be the most popular system being used in about 200,000 libraries of 135 countries. In USA 95% of the school library 25% of the academic libraries, and 20% of the special libraries use it. It has been translated in about 30 languages including Hindi Arabic, Chinese, French, German, Greek, Hebrew, Icelandic, Italian, Korean, Norwegian French, Spanish and Vietnamese. DDC is also used by various trade bibliographies, 62 national bibliographies, including the Indian National Bibliography. DDC numbers are available on MARC tapes, and in CIP data. Newly opened libraries mostly prefer to use the DDC for arranging their books. Continuous revision, strong institutional support from the Library of Congress at the beginning, now from Online Computer Library Center (OCLC) make DDC an easy choice acceptable to many professionals.

Self Check Exercises

1) Name the salient features of the DDC.

2) What are the advantages of the WebDewey?

Note: i) Write your answers in the space given below.
ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

.....

3.3 UNIVERSAL DECIMAL CLASSIFICATION (UDC) SCHEME

The Universal Decimal Classification (UDC) is a worldwide popular general classification scheme covering all fields of knowledge. It is an analytico-synthetic classification scheme, where the concepts can be broken down into simple

concepts (analysed) and then combined (synthesised). The UDC can be applied in different contexts that make it more flexible than other general classification schemes. It is considered as a sophisticated indexing and retrieval tool for documentation and information services.

3.3.1 Historical Background

The UDC is the brainchild of the two Belgians, Paul Otlet and the Nobel laureate Henry La Fontaine, who began working on a publication called '*Repertoire Bibliographique Universel*' in 1885. Otlet and La Fontaine built their system on the foundation of Melvil Dewey's Dewey Decimal Classification (DDC). Melvil Dewey conceived his scheme to be applied to the arrangement of books on shelves, whereas Otlet and La Fontaine conceived their scheme to be applied for the retrieval of documents as well as for arrangement of books on shelves. The first complete edition of UDC was published between 1905 and 1907 in French language. They established *Institut Internationale de Bibliographie* (IIB) in Brussels, which was the first publisher of UDC. The IIB later became *International Federation for Documentation* (FID) in 1937 and moved to The Hague. Subsequently, FID changed its name to *International Federation for Information and Documentation*. The FID was the centre for the management and maintenance of UDC till the formation of UDC Consortium in 1992. The official working languages of the UDC are French, German and English since its inception. The UDC is published in different editions, like, complete edition, medium edition, abridged edition and web-based online edition.

3.3.2 The UDC Consortium

Presently UDC is owned by the UDC Consortium (UDCC) since January 1992. The UDCC maintains multilingual *Master Reference File* (MRF), which is an electronic form of UDC schedules with all data held in a database. The UDCC publishes an annual journal, named '*Extensions and Corrections to the UDC*'. This journal contains reports from users, articles and revision proposals as well as approved amendments and revisions. The UDCC also maintains a website for general information and information on revision.

3.3.3 English Editions of UDC

British Standard Institution (BSI) started publishing UDC English editions as *BS 1000* series of documents. The BSI now publishes UDC in two formats: printed and online. The different English language editions of UDC, published by BSI, are:

i) The UDC Complete Edition

The UDC Complete Edition, published in 2005, contains over 65,000 entries and it is available in two volumes: Volume 1 – Systematic Tables; and Volume 2 – Alphabetical Index. Volume 1 includes main tables and auxiliary tables.

ii) The UDC Medium Edition

The UDC Medium Edition (BS1000M) was published in 1985 and revised in 1993, containing over 40,000 entries. It is available in two volumes: Volume 1 – Systematic Tables; and Volume 2 – Alphabetical Index. Volume 1 includes main tables and auxiliary tables.

iii) The UDC Abridged Edition

The latest UDC abridged edition, published in 2003, contains over 4,100 entries. It was formerly known as the UDC pocket edition.

iv) **UDC Online**

The UDC Online is the electronic version of UDC Complete Edition. This is subscription based and has additional functionality.

3.3.4 Features and Structure of the UDC

The UDC is a hierarchical and systematic scheme for classifying documents covering the whole range of recorded knowledge. The main tables consist of enumerated schedules for simple or basic subjects. The UDC has many features which are innovative and unique in nature. It has ability to express not just simple subjects but also compound subjects showing relationships between them. Like DDC, knowledge is divided into ten classes. Each class is subdivided systematically into further divisions, with each subdivision further subdivided. The more detailed the subdivision, the longer the number that represents the class. Like the DDC, this is made possible by decimal notation. The introduction of auxiliaries – common and special – is another feature of UDC. The auxiliaries of the UDC permit the construction of compound numbers through synthesis of subjects or facets.

The UDC is an aspect classification. Here phenomena are subordinated to the aspect from which they have been taken. A phenomenon may occur in more than one class, for example, computers in manufacturing, education, designing, business processes, etc.

In the UDC all recorded knowledge is treated as a coherent system, which is built of related parts. This is in contrast to a specialised classification in which related subjects are treated as subsidiary eventhough they may be of major importance in their own right.

The principles of facet analysis, though they are not explicit, are inherent in the structure of the UDC.

3.3.4.1 Notation

UDC uses a notation of mostly Indo-Arabic numbers used decimally, supplemented by a few other signs and symbols. The symbols chosen for UDC notation are non-language-dependent, and universally identifiable. Initially, UDC had a few symbols for auxiliaries; in later editions some symbols have been added to facilitate representation of more complex or compound subjects.

3.3.4.2 Structure of UDC (BS1000M: 1993)

The Volume 1 of UDC medium edition consists of systematic tables, i.e. main tables and auxiliary tables. The Volume 2 of UDC medium edition consists of alphabetical index.

Main Tables

The main tables of UDC comprise of ten main classes that represent whole of universe of recorded knowledge. But presently nine main classes are available, whereas one main class is kept vacant for accommodating new subjects in future. These main classes are:

0 Generalities
1 Philosophy. Psychology
2 Religion. Theology
3 Social Sciences
4 (kept vacant at present)
5 Mathematics and Natural Sciences
6 Applied Sciences

7 Fine arts. Applied arts. Entertainment. Games. Sports
8 Language. Linguistics. Literature
9 Archeology. Geography. Biography. History

Each main class is subdivided into its logical parts, with each subdivision further subdivided. The more detailed the subdivision, the longer the number that represents it. For ease of reading, long notational elements are broken up into 3-digit units by means of the point (.), which has no other significance (e.g., 123.4 or 123.456.7 and so on). But unlike DDC, UDC does not use 0 (zero) in the last digits for the first two levels of main classes (e.g. 5, 51; instead of 500, 510 as followed in DDC). An example is shown below to give you an idea on the hierarchical structure of UDC, as described in the Main Tables:

3	Social Sciences
37	Education. Teaching. Training. Leisure
378	Higher Education. Universities. Academic Study
378.1	Organisation of higher education
378.11	Organs and management of universities, etc.
378.112	Governing bodies. Council. Senate

This way the classification number is generated from the main class to the narrower subjects, and then the isolate ideas are combined using auxiliaries.

Auxiliary Tables

The auxiliaries in UDC are of two kinds: common and special. The common auxiliaries refer to generally recurrent characteristics, like, time, language, etc. The special auxiliary subdivisions refer to locally recurrent characteristics.

- ***Common Auxiliary Tables***

The common auxiliary tables provide notation for relationships or recurring concepts. The common auxiliary subdivision consists of numeric tables, in which concepts are enumerated and arranged hierarchically. The common auxiliary tables use different symbols to incorporate different concepts or ideas or facets. The common auxiliary tables, according to the UDC Medium Edition 1993, along with their applications are discussed below:

Table Ia Co-ordination. Addition

In Ia auxiliaries, '+' (plus) and ' / ' (stroke) signs are used to denote coordination or addition of classes.

The '+' (plus) sign connects two or more separated, but non-consecutive, UDC numbers to denote a compound subject for which no single number exists. Example:

(540+510)	India and China
51+53	Mathematics and Physics

The ' / ' (stroke) sign connects first and last of a consecutive UDC numbers to denote a broad subject, or a range of concepts, which may be called a consecutive extension. Example:

(4/6)	A group of Europe, Asia and Africa
592/599	Systematic Zoology (class 592 to 5999)

Table Ib Relation. Sub-grouping. Order-fixing

Ib auxiliaries use ':' (colon), '[ ]' (square brackets) and '::' (double colon) signs to denote relation, sub-grouping or order-fixing.

The ':' (colon) sign indicates relationship between two or more subjects by connecting their UDC numbers. Unlike the '+' (plus) and ' / ' (stroke), ':' (colon) restricts relation rather than extends the subjects it connects. The notation on either side can be reversed, like, A:B can be expressed as B:A. Example:

316.64:342 Social view in public law

If necessary, this can be expressed as 342:316.64. Similarly,

7:17 Art in relation to ethics

The '::' (double colon) sign indicates relationship between two or more subjects by connecting their UDC numbers, but it used to fix the order of the component numbers of a compound subject. The notation on either side cannot be reversed, like, A::B cannot be expressed as B::A. Example:

77.044::796 Sports photography

The '[ ]' (square brackets) sign may be used as a sub-grouping device within a complex compilation of UDC numbers, in order to clarify the relationship of the components. Sub-grouping may be required when a subject denoted by two or more UDC numbers linked by '+', ' / ' or ' : ' signs as a whole. Example:

783:[283/289] Protestant Church Music

Table Ic Common Auxiliaries of Language

The Table Ic enumerates the notations of languages. The = (equal sign) is used before the language notation to denote a specific language. Example:

=214.21	Hindi
51=214.21	Mathematics in Hindi language

Table Id Common Auxiliaries of Form

The Table Id enumerates the notations of documentary forms. The (0..) is used along with the form notation to denote a specific documentary form. Example:

(031)	Encyclopedia
5/6(031)	Encyclopedia of Science and Technology

Table Ie Common Auxiliaries of Place

The Table Ie enumerates the notations of places. The (1/9) is used along with the place notation to denote a specific place. Example:

(1-11)	Eastern
(261.268)	English Channel
(541.23)	West Bengal
37(541.23)	Education in West Bengal

Table If Common Auxiliaries of Ethnic Grouping and Nationality

The Table If enumerates the notations of ethnic groups and nationality. The (=...) sign is used along with the notation to denote a specific ethnic group or nationality. Example:

(=1.540)	Indian
(=1.4)	European
(=1.253)	Forest dwellers

Table Ig Common Auxiliaries of Time

The Table Ig enumerates the notations of date, point of time or range of time. The "..." sign is used along with the notation to denote a specific date or time. Example:

"20"	Twenty first century
5"20"	Science in the twenty first century
"3453"	Evening

Table Ih Specification by Non-UDC Notation (e.g. 1/9, A/Z)

Ih auxiliaries use sign * along with either 1/9 numeric or A/Z alphabetic notation from non-UDC sources to denote specific concept or thing. Example:

523.44*433	Minor planet Eros (IAU authorised number for Eros is 433)
625.711(540)*NH8	Road engineering with special reference to Indian national highway number 8

Table Ii Common Auxiliaries of Point-of-view

The Table Ii enumerates the notations for point of views. The (.00..) sign is used along with the notation to denote a specific point of view for a subject or concept. This auxiliary has a systematic table that is spelt out in detail. Example:

.003	Economic point of view
6.003	Economic point of view of Technologies and Applied Sciences
.002.6	Product point of view
688.322.002.6	Digital computer as a product

Table Ik Common Auxiliaries of Materials, Persons and Personal Characteristics

The Table Ik enumerates the notations for properties, materials, persons and personal characteristics. The –03 sign auxiliaries denote the materials of an object or product or thing. This auxiliary has a systematic table that is spelt out in detail. Example:

-033.52	Crystal glass
730-033.52	Sculpture using crystal glass
-035.54	Furs

The –05 auxiliaries denote the persons concerned and personal characteristics. This auxiliary has a systematic table that is spelt out in detail. Example:

-058.856	Foster parents
-057.81	Illiterates
730-057.81	Sculpture, made by illiterate persons

Table 3.1: Symbols of Common Auxiliaries at a Glance

Symbol	Used for	Associated with
+	Coordination / Addition	Table Ia
/	Consecutive extension	Table Ia
:	Simple Relation (May be reversed)	Table Ib
::	Order Fixing (Cannot be reversed)	Table Ib
[...]	Sub-grouping	Table Ib
=	Language	Table Ic
(0...)	Form	Table Id
(1/9)	Place	Table Ie
(=...)	Ethnic Grouping and Nationality	Table If
"..."	Time	Table Ig
*	Notation from non-UDC sources	Table Ih
A/Z	Direct alphabetical specification	Table Ih
.00	Point of view	Table Ii
-03	Materials	Table Ik
-05	Persons and personal characteristics	Table Ik

- ***Special Auxiliary Subdivisions***

Special auxiliary subdivisions use –1/–9, .01/.09 and '0/'9. These are limited in their scope. Each series is used to denote recurrent concepts in that part of the main tables for which it is designed and scheduled or in certain other sections where specially indicated in the schedule. Special auxiliaries are always schedule dependent, and cannot be applied to other classes unless specified.

The special auxiliary subdivisions use three kinds of notation

i) **The Hyphen series:** –1/–9 (i.e. –1 to –9) are mainly analytical or differentiate in function. These indicate elements, components, properties and other details of the subject denoted by the main number to which they serve.

Example: 82–1/–9 English Literary forms, genres

82–1 English Poetry. Poems. Verse

ii) **The Point-Nought series:** .01/.09 (i.e. .01 to .09) provides sets and subsets of recurrent concepts, like, aspect studies, processes, operations, plant and equipments.

Example: 528 Geodesy. Surveying. Photogrammetry. Cartography

528.01 Preparatory work. Making of stations. Signal construction

iii) **The Apostrophe series:** '0/'9 (i.e. '0 to '9) are used for more specific instances than –1/–9 and .01/:09. These are most of the times synthetic or integrative in function and denote compound subjects by compound notation.

Example: 669 Metallurgy

669.35'5'6 Copper-zinc-tin alloys

(derived from 669.35, 669.5 and 669.6)

- **Parallel Subdivisions**

When one portion of the classification is divided like another, this is indicated by the symbol ≈.

Example:

611.11/.14 ≈616.11/.14 Principal divisions of *Anatomy* are similar to principal divisions of *Special Pathology*

- **See Also References**

Reference symbol right arrow (→) is used in UDC to draw attention of related aspects of a given subject, i.e. this symbol indicates *See also* reference.

Example:

55→502.6 Earth Sciences

See also Landscape

3.3.5 Filing Order

The filing order is required for the purposes such as filling or the organisation of books on the shelf in a library context, or for creating a systematic display in an OPAC or for the preparation of bibliographies by subject. The filing order of UDC symbols is based on a progression from the general to the specific. The independent auxiliaries come first. Then aggregation of several numbers (using Ia auxiliaries) comes before the simple component number. Then comes the numbers with auxiliaries as suffixes. The Table 3.2 shows the filing order of both simple and compound numbers.

Table 3.2: Filing Order

	Symbol	Examples	
Independent auxiliaries	=	=211	Sanskrit Language
	(0...)	(0.035.23)	Microfiche
	(1/9)	(540)	India
	(=...)	(=1.540)	Indian
	"..."	"20"	21st Century
	+	622.341.1+669.1	Iron ore mining and ferrous metallurgy
	/	622.341.1/.2	Mining of iron ores and manganese ores
	Simple Number	622.341.1	Iron ore mining
Auxiliaries as suffixes	:	622.341.1:338.58	Economic costs of iron ore mining
	::	622.341.1::338.58	Economic costs of iron ore mining
	=	622.341.1=214.21	Documents in Hindi on iron ore mining
	(0...)	622.341.1(035)	Handbook of iron ore mining
	(1/9)	622.341.1(1-925.36)	Iron ore mining in the Himalayas
	(=...)	622.341.1(=1.1.253)	Iron ore mining amongst forest-dwellers
	"..."	622.341.1"6387"	Iron ore mining in the late Iron Age
	*	622.341.1*Fe_2O_3	Iron ore mining of red hematite (Fe_2O_3)
	A/Z	622.341.1GEO	Goethite iron ore mining
	.00	622.341.1.002.68	Scraps from iron ore mining
	-0	622.341.1-057.21	Skilled workers in iron ore mining
	-1/-9	622.341.1-78	Protective devises and measures in iron ore mining
	.0	622.341.1.03	Geological character in iron ore deposits
	'	622.341.1'17	Tailing, waste and residue from iron ore mining
	Next Simple Number	622.341.11	Magnetite-hematite ore mining

3.3.6 Citation Order

The citation order is governed by the rules and conventions of any classification scheme for such purposes as displaying the concepts in the UDC schedules. The arrangement most commonly used is known as standard citation order. The standard citation order offers a framework, i.e.:

Thing – Kind – Part – Material – Property – Process – Operation – Agent – Space – Time

The UDC follows this framework, but it has provision to adapt the citation order to fit in with local requirements.

If you look at 524 schedule you will notice the citation order, as depicted below:

Things	524	Universe. Stars, etc.
Kinds	524.31.01	Super giant stars
	524.31.02	Giant stars
	524.31.03	Subgiant stars, etc.
Parts	524.6	The Galaxy
	524.63	Local region the Galaxy
	524.64	Galactic centre, etc.
Materials	*Not applicable*	
Properties	-3 subdivisions	Properties and phenomena
Processes	-4 subdivisions	Processes of bodies and systems
Operations	-1 subdivisions	Mode of treatment. Working methods, etc.
Agents	520	Astronomical instruments and techniques

Many types of concept can be identified within the standard citation order, as shown above. The elements, which are not appropriate to the discipline can be unavailable or absent.

3.3.7 Qualities of the UDC Scheme

The UDC is an analytico-synthetic classification scheme which is popular worldwide and used in libraries, museums, archives, bibliographic databases, information services and Internet. It is accepted worldwide for many reasons and qualities. These are summarised below:

i) The UDC is very flexible in nature.

ii) The UDC like the DDC has been published in Complete, Medium, Abridged, and Web formats, which can be used by libraries according to the requirements.

iii) The UDC performs well to applications in other languages and scripts. Its notation overcomes all language barriers.

iv) Due to its versatility it can be utilised in multiple fields including libraries, museums, archives, in documentation and Internet, not only for shelf arrangement or collection display, but also for information retrieval or as a metadata.

v) The UDC is updated regularly as and when the new subjects or ideas are coming up. The UDCC brings out an annual journal that informs us about the new classes, revisions, and cancellation of any item.

vi) The UDC handles simple subjects, compound subjects, complex subjects and isolate ideas quite efficiently. The combination of isolate ideas with a subject is very systematic and flexible.

vii) The contextual analysis of subjects is also possible in UDC that provides flexible treatment of documents based on individual institution's requirements.

viii) It is easier to manipulate the UDC to accommodate advances in knowledge because of greater scope for creating new synthesised numbers for concepts or simply inserting a new number as required without the need to reach general editorial agreement.

ix) Due to its abbreviated nature and vocabulary it is easily updated and enables worldwide standardised indexing.

x) Within the UDC many concepts that look new are in reality a combination of existing ones and so can be immediately expressed.

xi) Due to the UDC's incredibly flexible character it provides interface to conversion in a digital computer format.

xii) The UDCC maintained Master Reference File is very comprehensive in nature and updated at regular interval.

The analytico-synthetic nature of UDC scheme makes it very flexible and accommodating. Although this scheme is not fully faceted, it has very powerful facility, which permits the joining of any part of the classification with any other. UDC can be applied in different contexts and in different applications in special libraries and information centres. It is now being vigorously geared for use in online catalogues and websites on the World Wide Web.

Self Check Exercise

3) Name the different editions of UDC in English.

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

.....

.....

.....

.....

.....

3.4 LIBRARY OF CONGRESS CLASSIFICATION (LCC) SCHEME

As obvious from the name, this system is developed (from 1904 onwards) and used by the Library of Congress, Washington, DC. The Library of Congress, *de facto* the US national library, is the largest library of the world in terms of collection and services. It has now more than 90 million items of reading material. The library was shifted to its present building in 1897. Then the Librarian Herbert Putnam (1861-1955) after having decided to use a new system of its own for L.C. assigned the job in 1898 to Charles Martel (1860-1945) and JCM Hanson (1864-1943). This scheme was based somewhat on C A Cutter's Expansive Classification (1893). It consisted of 21 classes in 29 parts, in 48 volumes running to about 11,000 pages. The system comprises a series of coordinated schedules.

Each major class is virtually independent with its own form and geographical divisions and index. The system is purely enumerative, i.e. the class numbers are available ready-made and there is little provision of synthesis. As its another major characteristic, it is based on the literary warrant. Classes are simply based on and developed from what is available in the Library of Congress. It is not any theoretical mapping of the universe of knowledge. It is also available in CD-ROM entitled ***Classification Plus***. The CD version combines schedules with Library of Congress Subject Headings (LCSHs) and provides many more access points. It is easy to use and update.

3.4.1 The System

The latest edition is the 6th edition (1990) :

Library of Congress. Subject Cataloging Division

Processing Services: LC Classification Outline.

6th ed. Washington, DC: LC, 1990. ISBN 0-8444-0684-8

The main classes are:

A Generalia

B Philosophy

C/F,G History and Geography

H/L Social Sciences

M/P Humanities; P Languages and Literature

Q/T Science, Technology and Medicine

U/V Military Science

Z Library and Information Science

Letters I,O,W,X and Y are still vacant. Each of the main class is further divided alphabetically, e.g.,

QA	Mathematics	QD	Chemistry
QB	Astronomy	QE	Geology
QC	Physics	QH	Biology

Each class is further subdivided by one to four arithmetical numerals :

QA	1-939	Mathematics
	9-10.3	Mathematical logic
	75.5-76.95	Computer Science
	101-141.8	Elementary Mathematics
	150-272	Algebra

Decimal extensions are of recent use to accommodate new subjects. In the final shape the LC class number is bit more complex comprising of highly mixed notation. An LC class number is further extended by Cutter's author number and sometimes by year of publication; thus the class numbers are usually lengthy, e.g.,

HV 1481.S82 G34	Oldage Pension in Sweden by P.Gahrton.
PS 1331.H4.1974	Mark Twain by A Henderson (1974)

PR3629. A2L6	The Rape of the Lock, ed by Laugee and McHenry
QC16.E5	Albert Einstein –the physicist

3.4.2 The Use, Revision and Future

Though in design and details, it is not scientific, yet along with DDC and UDC, it is considered among the three 'big' library classifications. It is absolutely enumerative scheme so much so even form and space auxiliaries are enumerated. It is a popular scheme in USA; some other big libraries abroad are also its users. According to an estimate, it is being used by 60% of the research libraries in the USA. Its class number appear on MARC record and CIP data, and on some cooperative catalogues and commercial cataloguing services. It is also used outside the USA.

It is constantly revised, and has a great advantage of having both the designer and the main user as one body. The new changes made after a great deal of deliberations are announced in the quarterly bulletin: *LC Classification: Additions and Changes*. The CD-ROM version called *Classification Plus* includes LC subject headings and is available on annual subscription with quarterly updates. CD version is convenient and efficient as it allows search by keywords, class numbers, proximity options, and Boolean operations, etc. As a limitation not the entire system has yet been converted into CD-ROM format. It has an assured future mostly because of its institutional backing. It is a classic example of a highly successful enumerative classification system with scholarly details.

Self Check Exercise

4) State the salient features of the LC Classification.

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

.....

.....

.....

.....

.....

3.5 COLON CLASSIFICATION (CC) SCHEME

The CC projected as India's *de facto* national classification was conceived and developed during 1924/1932 and published first in 1933 by the Madras Library Association (MALA). Its author Dr. S.R. Ranganathan (1892-1972), called the Father of Indian library movement, was designated as the National Research Professor in Library Science in 1965 by the Government of India. His world famous system named Colon Classification is acknowledged as the greatest breakthrough in bibliographic classification since the DDC (1876). He developed an entirely new species of library classification called faceted classification. The CC is an analytico-synthetic freely-faceted scheme as contrasted from the enumerative classifications such as DDC (edition 1-15), the LC, and Rider's International Classification (RIC). The 4th edition (1952) incorporated many new and modern concepts and laid down the foundation of the modern version of the CC. The sixth edition published in 1960, and reissued with amendments in 1963 was the last edition produced by Ranganathan.

You have studied the CC 6th edition in detail in BLIS programme. Here we will be discussing the 7th edition and its features.

3.5.1 Features of 7th Edition of CC

The 7th edition (1987) was revised and edited by Professor M.A. Gopinath then the Head, Documentation Research and Training Centre, Bangalore (and Secretary, Sarada Ranganathan Endowment for Library Science).

Ranganathan, S.R.: Colon Classification, 7th ed,

Revised and edited by M.A. Gopinath. Bangalore: Sarada

Ranaganathan Endowment for Library Science, 1987. Volume 1, xiv, 332p.

So far only one volume containing schedules and common isolates, etc. has been published. The index volume is available only on a CD-ROM. In the new edition, the following 10 types of basic subjects have been identified.

1	**Main Basic Subjects**	2	**Non-main Basic Subjects**
1.1	Traditional	2.1	Canonical classes
1.2	Newly emerging	2.2	System constituents
1.3	Fused	2.3	Special constituents
1.4	Distilled	2.4	Environment constituents
1.5	Subject bundles		
1.6	Agglomerates		

In the new edition, the number of basic subjects has been enormously increased to about 750. However, the traditional basic subjects remain the same:

1/9	Newly emerging main classes.
A/M	Sciences
Δ	Mysticism and Spiritual Experience
N/S	Humanities
T/Z	Social Sciences

Some of the new basic subjects in CC are:

I	Universe of subjects	4	Mass communication
BT	Statistical calculus	LT	Physical exercises and Sports
M4	Smithy	PW6	Typewriting

Obviously, some trivial subjects have been raised to the status of basic subjects. In the 7th edition, the Common Isolates (CI) have made significant advances both in kind and number, mostly in Posteriorising Common Isolates (PCI). the new schedules of (CI) are:

Common Matter Property (Chap.DL, pp. 95-104)

Common Energy Isolates (Chap.DK, pp. 93-95)

The Anteriorising Common Isolates (ACI) are now added with the indicator digit " " (double inverted comma), e.g.,

Journal of Mathematics

B"m

Encyclopaedia of Chemistry

E" k

But most important of the additions is the enumeration of Environmental Divisions (ED) (Chap DD, pp-54-56) which can be added to any subject using "-9" as the indicator digit:

Un4 High altitude

Y311 Rural

High altitude engineering

D-9 Un4

Rural Economics

X-9Y311

Another important addition is the use of Spectators of Kind I and Kind II to be denoted respectively by "-" hyphen and equal to = signs:

Partially Sovereign State

W, 1-12

Dialects of English Language

P, 111=d

The Alphabetical Device (AD) is frequently used and the components of a multinomial term are joined by "+" plus:

Shakespeare's King Lear

O,111,2J64, K+L

Phase relation digit 0 (zero) has been replaced by & (ampersand), and a new kind of phase relation namely 'Tool' has been recognised:

Mathematics for engineers

B&bD

Mathematics as a tool to study science

A & eB

In the new edition, now it is obligatory to put a comma "," between the (MC) and first Personality (1P1), e.g.,

S,1 Child psychology (instead of S1)

P,15 Sanskrit Language (instead of P15)

As another fundamental change, now three varieties of the fundamental category 'Matter' have been discovered:

Matter - Property (M-P)

Matter - Method (M-M)

Matter - Material (M-Mt)

What was energy category in the CC-6, now, as a whole, has been redesignated as Matter- Property e.g.

	CC-6	CC-7
Library Classification	2:51	2;5

Human physiology	L:3	L;3
Constitutional history	V:2	V;2
Constitutional History of India	V44:2	V,44;2

3.5.2 Notation of CC-7

The following five new digits have been added in the notational system, while none of the earlier digits has been discarded: (except some of the Greek alphabets)

+ (Plus), = (equal to). " (double inverted comma).

* (asterisk), & (ampersand)

The notation in the CC-7 comprising of 74 digits (60 semantic and 14 indicator digits) has been divided into the following six species:

1.	A/Z (Roman Capitals)	26
2.	Δ (Greek Delta)	01
3.	0/9 Indo-Arabic numerals	10
4.	a/z (i,l,o excluded) Roman smalls)	23
5.	Indicator digits with anteriorising value * " ←	03
6.	Ordinary indicator digits () &' . :; , - = → +	11 = 74

(Asterisk, double inverted comma , backward arrow, parenthesis, ampersand, inverted comma, dot, colon, semi colon, comma, hyphen, equal to, forward arrow, and plus respectively)

The notational base of the CC is the widest ever in any classification system. On this count, the CC-7 is very hospitable, though it has made the notation and consequently the system quite complicated: Class number looks lengthy like algebraic equations even unwieldy. Ranganathan sacrificed the brevity and simplicity of notation to make it finely hospitable, and to produce co-extensive class numbers.

The new edition has many internal inconsistencies. Nevertheless, it is a useful contribution containing a number of innovations. Expert systems are being developed for automatic synthesis of CC numbers at the DRTC Bangalore. Index to CC-7 is available only in a CD-ROM form titled CINDEX. This database has been designed and developed by Sarada Ranganathan Endowment for Library Science using WINISIS 1.4 software developed by Unseco.

3.5.3 Revision, Use and Status

Despite being projected as India's national scheme of classification, it is not a widely used system in India. Though no register of its users has been maintained, but according to a very favourable estimate some 24% of the libraries are using this system in India. CC class numbers are also given as a bibliographic element in the entries of the Indian National Bibliography. There is no national committee or a substantive institutional backing to advise or to take responsibility of its revision or publication. It is now published by the Sarada Ranganathan Endowment for Library Science, Bangalore.

The system is faceted in nature and is based on postulates and principles integrated into a coherent and fully and finally developed theory of classification. The contribution of the CC lies in its facet analysis technique, the concept of fundamental categories and canons which are practical. It provides scientific guidelines for construction of any new classification system. The theory of the CC is in fact considered as the theory of classification in general which is taught in many library schools the world over. Many depth classification systems are being designed using the CC methods and theory. But survival of the system in its present body seems uncertain.

Self Check Exercise

5) Enumerate the various notational digits of the CC-7.

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

.....

.....

.....

.....

.....

3.6 BIBLIOGRAPHIC CLASSIFICATION (BC) SCHEME

The system was designed during 1910-1953 by Henry Evelyn Bliss (1870-1955), the then Librarian, City College (Now, City University), New York. Bliss occupies a very respectable place in the history of library classification as, to him goes the credit of basing his classification on a very sound theoretical foundations. Classification study and design was his life time occupation in which he earned distinctions. Before publication of his classification, Bliss authored two classical works on the Organisation of Knowledge in 1929 and 1933. The Bibliographic Classification (BC) was announced in 1910, and its outline was published in 1935 under the title "A system of Bibliographic Classification". The classification itself was published in four volumes in 1940, 1947, 1952 and 1953.

Bliss, H.E. *A Bibliographic Classification extended by systematic auxiliary schedules for composite specification and notation.* New York: H.W. Wilson, 1940-1953. 4 Vols.

It took much time in editing and production. By the time the last volume came out, the earlier volumes became outdated. The scheme could not find many users as it arrived quite late and by that time the systems like DDC, LC and UDC were well entrenched into the field. However, it did find a few admirers and users in the UK. H.E. Bliss is the proverbial prophet unsung in his own land.

3.6.1 The Second Edition, the BC-2

In 1967, Bliss Classification Association (BCA) was formed in the UK under the leadership of J. Mills. The aim of BCA was to receive, and revise the system and promote its use. It started work on the 2nd edition simply known as BC-2 edited by J. Mills.

Mills J, et al: Bliss Bibliographical Classification, 2nd ed. London: Bowker Saur, 1977+ (Earlier volumes were published by Butterworth, London).

BC-2 is in 19 parts of which many have been published. The work has been slow due to lack of financial support and manpower. Nevertheless, the BC-2 has many new features and merits of its own. It is also known as Mill's BC-2 which is a

tribute to the genius and dedicated efforts of J. Mills. The classification Research Group (CRG, London) played an active and vital role in its revision.

3.6.2 Special Features of BC

The BC has many unique features and merits:

a) The order of main classes is based upon what Bliss called the Educational and Scientific Consensus. He even claimed that he had found a permanent order of subjects in the universe of knowledge. The claim of discovering a permanent order cannot be admitted, as no such order exists permanently in the universe of knowledge. This order being social, changes from time to time and society to society. Nevertheless the BC order of main classes is commendable.

b) Bliss provided alternatives in the sequence, as e.g. Economic history can be placed either with economics or with history in general. The religion can either be placed at P or at Z.

c) Bliss believed in simple and short notation now comprising of Roman capitals and Indo-Arabic numerals. It usually violates hierarchy to have brief notation e.g.

U	Technology
UE	Engineering
UHC	Construction
UJ	Architecture

d) The scheme uses retroactive notation which can be combined backward, e.g.

2HL	Coloured material
2P	TV
2PHL	Coloured TV

e) It is the latest general classification scheme based upon modern facet analysis and helpful sequence methods expounded by Ranganathan. But it avoids many of Ranganathan's complexities and faults. Therefore, it has the best of both the worlds, and possesses many unique but commendable features. This scheme has bridged the gap between classification theory and practice.

3.6.3 Structure, Auxiliary Tables and Use

Its main classes are as follows:

A	Philosophy	K	Social Sciences
AL	Logic	KA	Sociology
A	Mathematics	L/O	History
		M	Modern History
B	Physics	P	Religion (Alternative Z)
C	Chemistry	Q	Social Welfare
D	Astronomy	R	Political Science
E	Biology	S	Law
F	Botany	T	Economics
G	Zoology	U	Technology
H	Man, Anthropology	V	Fine Arts

I	Psychology	W	Language and Literature
J	Education	Z	Religion (Alternative P)

In addition, it has six auxiliary tables:

1	Common sub-divisions	4	Languages
2	Persons	5	Ethnic Groups
3	Places	6	Periods of time

The scheme is very strong in main class J Education, and Q Social Welfare. The class J is in its second edition, and had influenced other classification systems especially 370 Education in the DDC-20 (1989). On this account, it is gainfully applied in special libraries, libraries of welfare departments and colleges of education. There are some 90 users in Europe, Australia and Africa. It is suitable for preparing thesauri, and a model for teaching faceted classification.

Each volume has its own index prepared according to chain indexing. A cumulative index will be provided on completion of the whole scheme. The BC-2 is constantly revised by J. Mills with the assistance of CRG (Classification Research Group) members. Revision is sound but slow. The scheme is brilliant and hopes to find its deserved place soon.

Self Check Exercise

6) Enumerate the salient characteristics of the BC.

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

.....

.....

.....

.....

.....

3.7 LIBRARY BIBLIOGRAPHICAL CLASSIFICATION (BBK/LBC) SCHEME

Every classification naturally reflects the culture of time and society where it is designed. Almost all the major classification systems are biased towards Christian and Capitalistic subjects and have been averse to Marxism and Leninism: a reflection of the cold war days. Thus, libraries in socialistic countries found most of the classification systems biased and useless for them. It prompted the then Lenin State Library at Moscow to design a system of its own based upon the Marxist ideology. The Lenin State Library with 40 million collection was the national library of the erstwhile USSR. It is now renamed Russian State library. The decision to create a Soviet classification system was taken back in 1920s though work on it started quite late.

3.7.1 Structure of the BBK

Its name translated in English is "Library-Bibliographical Classification". In Russian its abbreviation is BBK. It was designed by a broad based team of experts under the editorship of O.P. Teslenko. The BBK was first completed in 30 volumes, and lists 34,000 main and 51,000 compound numbers indicated by 28 letters of Russian (Cyrillic) alphabet.

It is the first classification which provides alternative set of notation namely either 28 letters of Cyrillic alphabets or Indo-Arabic numerals:

1	Interdisciplinary knowledge.
2/5	Science, Technology. Agriculture and Medical sciences
6	Social Sciences
7	Culture and education
8	Language and literatures.
86	Religion, Atheism
87/88	Psychology
9	Universal subjects

Notation in general is highly mixed, hierarchical, faceted and synthetic. The punctuation marks are Point, Colon, Hyphen Dash, Quotation marks, Round Brackets, Equals, Stroke and Asterisk, e.g., T2 (4f/PA) History of France

The system is semi-faceted but detailed, and it based upon the literary warrant of books and journal literature available in Russian libraries of Moscow and Leningrad, now St. Peter Square . The main tables are supported by a large number of auxiliary tables both general and special, mostly modeled on UDC. The class numbers are usually lengthy. Each class has its own index derived by chain indexing. As an estimate about 280,000 libraries and information centres in the former USSR republics are using it. It was also used in the erstwhile East Germany, Bulgaria, Vietnam and North Korea. In addition to the main tables, 5 volume abridgement (1970-1972), one volume abridgement (1975) for public libraries another one volume (1978) for school libraries and 4 volumes (1980-1983) for regional libraries are available. The system is constantly revised, and LBC-2 is under preparation; and the system is being adopted to the needs of online catalogues. Though its future at present seems uncertain, yet it is hoped that it will continue to be used in Russia.

Self Check Exercises

7) Why was BBK designed?

8) State its main classes.

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

.....

.....

.....

.....

.....

.....

3.8 BROAD SYSTEM OF ORDERING (BSO) SCHEME

Apart from the UDC, Broad System of Ordering is only another classification originally sponsored by an international organisation, in this case UNESCO with the technical expertise provided by the FID/BSO panel. UNESCO wanted in 1971 a switching language for the UNISIST programme (Universal Information System in Science and Technology). A switching language means a link or intermediate or conversion language from one indexing language to another.

Therefore, it is not a classification system in the ordinary sense but a tool for information exchange for switching between classifications or information retrieval systems.

3.8.1 Structure

The BSO panel working mostly under Eric J Coates and G.A. Lloyd completed the first draft in 1975. Now the following edition is finally available:

> Coates, Eric J, et al: *Board System of Ordering* (BSO): Three 3.5 inch disks in MS-DOS format. Hearts, UK: BSO Panel.

The major subject areas are outlined below:

100	Knowledge in General	500	Humanities and Social Sciences
200	Science and Technology (Together)	600/890	Technology
300	Life Science	910	Language and Literature
460	Education	940	Arts
480	Sports and Games	970	Religion

It has about 6800 very broad subdivisions in all and two schedules of common facets for time and Space. It uses pure notation of numerals used mellesimally (dividing by 1000) and centesimally (dividing by 100) in the pattern of 3,2,3. Only two punctuation marks are used '–' (hyphen) and ' , ' (comma). The scheme is faceted and has provision for synthesis based on modern theories of facet analysis. It uses institutional warrant rather than literary warrant. It is useful for conventional disciplines, multidisciplinary as well as mission oriented topics.

360,20	Agricultural biology
395,60	Environmental pollution
395,60-230	Chemical pollution where 230 stands for chemicals.
360-580	Agricultural economics

Thus '-' hyphen serves as a phase relation symbol. It also uses A/Z notation sparsely.

3.8.2 Use

The scheme is only available in three floppies of 3½ size. Ironically the scheme is not being used as a switching language for which it was originally intended. However, a few libraries are using it for arranging books on the shelves. No further work is being done on it now. However, the system deserves to be considered as a contribution to the realm of intellectual organisation of knowledge for its order of main classes and synthesis of topics with simple notation.

3.9 SPECIAL CLASSIFICATION SYSTEMS

In the previous sections, you have learnt about the general classification systems which deal with universe of knowledge. But many information professionals working with special libraries and information centres face several problems and difficulties in classification of the micro-subjects or inter-disciplinary subjects. General classification systems organise the knowledge at a certain level or at a certain details. The new developments in knowledge are relatively undiscovered in general classification systems. At the same time, forms of documents are also very important to some specialised institutions. Thus, general classification systems

may not be able to cope up with the requirements of special libraries that hold specialised forms of documents, like patents, standards, maps, documentary films, feature films, photographic slides, transparencies, audio materials or audio-visual materials and also the specialised areas in the knowledge front. As a result, some special classification schemes are evolved to organise documents in specific subject areas or in specific forms. Rita Marcella and Robert Newton [1994] define the special classification scheme as:

> "A special classification scheme is a system with a narrower focus. Basically, special schemes attempt to do one of the three things: they either attempt to provide a classification for a narrower subject, a discipline of study or an area of professional practice; or they provide a scheme which is aimed at a particular user group; or they cater for a form of material which requires special treatment."

Dr. S. R. Ranganathan [1967] defined the special classification scheme as "scheme designed for depth classification of micro subjects, going only with one and only one specialised subject field".

3.9.1 Benefits

Some of the benefits of special systems of classification are:

i) Special systems can provide enough and desirable details for dealing micro subjects.

ii) Special classification systems can provide less-lengthy class numbers for compound and micro subjects, as compared to most general classification systems.

iii) Special systems are able to meet the special requirements of a particular special library or information centre.

iv) Special systems mostly have flexibility to accommodate newly emerging subject fields.

v) Special systems can accommodate all the aspects of the subject to be kept together, whereas a general system may scatter them.

vi) Special systems can keep together a specific form of materials and scattering of documents of such form can be avoided.

vii) The special systems can be up to date. The updation of special systems can be easier as the subjects grow.

viii) Inter-disciplinary subjects are scattered in general systems whereas special systems can accommodate the subjects together according to the needs of the special libraries and their clientele.

3.9.2 Scope

Scope of special classification system is very wide with varying degrees. Special classification system can be restricted to specialised areas such as:

- Restricted to specialised subject areas, e.g., Astrophysics, Biotechnology, Biochemistry, Computer Science, Western Music, Indian Culture, Information Management, Vetenary Science, and Indian Literature.
- Restricted to an association of topics that are generally scattered in general classification schemes, e.g., Indology, information and communication technology, journalism and mass communication, non-conventional energy sources, marine science and oceanography.

- Restricted to certain physical forms, e.g., still photographs, maps, DVDs, audio-video materials, cartoons, paintings and drawings.
- Restricted to certain form of publications, e.g., patents, standards, trade catalogues, websites, and prospectus. Also some computer document formats of publications may be restricted like presentation (pps or ppt), textual (pdf or doc format) image (jpeg, bmp, gif, tiff, etc.), moving image and sound (mpeg, wav).
- Restricted to certain form of thought presentation, e.g., Non-fiction, fiction, Poetry, Drama, cartoon strips.
- Restricted to certain type of readers, e.g., children, students, neo-literates, visually challenged, physically challenged, travelers.

3.9.3 Approaches

In the design and development of special systems we can broadly see three types of special systems. These are:

i) **Fully Autonomous Special System**

Fully autonomous special system is designed independent of a general classification system to organise documents, which are confined to specialised subject fields or special formats or specialised user groups.

ii) **Partial Autonomous Special System**

It is designed autonomously for a specialised subject field, but is dependent upon general system for allied subject fields. This approach takes the advantage of a general system for allied subject fields, and it is expected to incorporate changes made in the general system in the light of expanding universe of knowledge.

iii) **Special System Dependent upon a General System**

According to Dr. S. R. Ranganathan, a special system can be designed in such a way that it is dependent on a general system. This means that a special system is nothing but an extension of a general system.

He emphasised the need for 'do all classification' that can meet all the requirements of the special system in different subject fields irrespective of the scope of subject (extension) and the depth of subject (intension). Dr. Ranganathan also suggested that a 'do all classification' should be truly faceted based on dynamic theory of library classification.

3.9.4 Some Examples

- The London Classification of Business Studies. Published in 1970, was designed upon anlytico-synthetic principles.
- Harvard Classification of Business Studies. Published in 1960, was designed upon enumerative principles.
- INSPEC Classification a classification scheme for physics, electro technology, computers and control. Published in 1988, was designed upon enumerative principles.
- The CI SfB system. For use in project information and related general information. Published in 1976, took a fully faceted approach and focused on construction industry.
- Moy's Classification Scheme for Law Books. First edition published in 1968 and 2nd edition in 1982. It used partial enumerative and partial synthetic principles.
- FIAF Classification Scheme for Literature on Film and Television. First edition published in 1980 and 2nd edition in 1992. It is primarily enumerative in structure. It also allows for some synthesis via the provision of Auxiliary Tables for Form. It is developed by International Federation of Film Archives.

- The Cheltenham Classification. A Library Classification for Schools. First edition published in 1937 and 2nd edition in 1958. It was designed upon enumerative principles.
- Early Stages Library Classification Scheme and Index. Published in 1992. It targets specifically primary school years. It is opted for a colour-coding approach along with notation.
- Iconclass: An Iconographic Classification System. First edition published in 1947 and revised edition in 1983. It is a non-bibliographic classification system in order to provide a means of identifying the content of works of western art in terms of their subject content, themes and motifs.
- ACM's Computing Classification Scheme (CCS).
- Ei Classification. It was developed by Engineering Information Inc. It is used at the Edinburgh Engineering Virtual Library.
- National Library of Medicine (NLM) Classification. It specialises in providing access to biomedical information.
- International Patent Classification (IPC). It is developed and maintained by World Intellectual Property Organization (WIPO).

The Documentation Research and Training Centre (DRTC), Bangalore and Classification Research Group (CRG), London have made substantial contribution in the development of special classification schemes and also development of methodology. The methodology developed at DRTC is based on Ranganathan's faced principles. The schemes developed at CRG are greatly influenced by Ranganathan's ideas.

The summary of methodology of construction of special classification scheme is: understanding the subject, identifying the users needs, collections of terms, analysis into facets, establishment of hierarchical relationships, preparation of draft structure of schedules, allocation of notation, preparation of draft schedules, preparation of draft's index, assigning class numbers to sample of subjects, seeking opinion of the specialists on helpfulness or otherwise of the notations, finalisation of schedules after scrutiny and preparation of set of rules for classifying.

The DRTC in collaboration with Sarada Ranganathan Endowment for Library and Information Science has developed a number of depth schedules based on Colon Classification on various subject fields such as Production Engineering, Gear, Locomotives, Screw, Nut, Bearings, Combustion Engines, Banking, Library Science, Food Technology, Solar Energy and some others. The DRTC has made major contributions in classification research, especially in depth classification and special systems since its establishment in 1962.

The CRG has also made major contributions in design and development of special classification schemes sins its inception in 1952. Some of such special schemes are: Classification for Social Sciences (by B.F. Kyle), Scheme for Diploma Technology (by J.E.L. Farradane), Scheme for Occupational Safety and Health (by D.J. Foskett), Scheme for Soil Earth Science (by B.C. Vickery), British Catalogue of Music Classification (by E.J. Coates) and Scheme for Library and Information Science Abstracts (LISA).

3.10 TRENDS OF RESEARCH IN CLASSIFICATION SYSTEMS

In the early 1980s the classification studies were in a low key. Use of computers in library and information work had something to contribute to that dismal situation. Information in a computerised database easily provides so many access points

that a user thinks the information has been arranged in a way particularly suitable to his/her requirements. Keywords searching was considered convenient and adequate. Hence, no need was felt using any classification system. Some computer enthusiasts even began to advocate the removal of classification studies from the library science curricula. However, in the late 1980s it was discovered that classification has many uses in online catalogues and can increase precision and recall ratios in computer aided information retrieval. As a result, at present the most dominant trend in classification studies is the use of classification in online public access catalogues (OPACs). Experiments have been conducted on the use of DDC, UDC, LC and BBK in online catalogues with a view to find out their effectiveness. UDC, DDC, LC and BSO are now available in electronic media. Computers are also being used as aids in revision of big systems such as the DDC, and most of the systems are now available in electronic format. Above all Internet search engines use classification, however simple, in retrieving information from the Web. Other trends that could be perceived are to:

1) base classification systems on the modes of growth of subjects;
2) use Standard and current terminology;
3) design general systems of classification to serve simultaneously as special and general classification systems. Most of the general classification systems are officially available in different versions of varying details;
4) make classification system free of national and cultural bias for use in other countries and cultures;
5) discover more varieties of auxiliary subdivisions;
6) make the notation more versatile and simple;
7) use classification and its principle in designing index languages and tools for vocabulary control;
8) design expert systems for automatic synthesis of class number by computers; and
9) experiment with Automatic classification by cluster analysis and work frequency count.

A healthy trend is of cooperation between classification systems and organisations instead of any rivalry or push back competition. The trend noticed is the cooperation among them for classification system designing. UDC and BC-2 are joining hands to make UDC scientifically faceted. UDC 61 Medicine has already been structured on lines of BC-2 as a result of active cooperation between the two. Similarly UDC and DDC are producing a joint Area Table. Editors of other classification systems are opted on the editorial boards of other rival systems.

It is interesting to note that all these trends are set in context of faceted schemes of classification. For a long time librarians have started feeling the need for a new general classification system which could reflect the modern structure of knowledge and is based upon the theory of facet analysis. Some scholars feel that an opportunity of designing a new system was missed in 1960s when a number of so called redbrick universities were established in the UK. Another such opportunity was not availed when British and French national libraries recently moved to their new magnificent and grand buildings.

In the first conference of the International Society for Knowledge Organization (ISKO) held in Germany in August 1990 late Eric de Grolier stressed the need for a unified classification system and its general principles. This is a deterring task because of its magnitude and the astronomical amount of money it requires. Development of new classification systems is also hampered by the fact that librarians, for many good reasons, prefer to continue with old and imperfect

systems. A classification once adopted in a library cannot be easily replaced by another, however, efficient and scientific the new system may be. Therefore, no new general classification is in the making though many specialised subject classification schemes are frequently designed. Librarians have learnt to live with the imperfections of general classification systems, and try to remedy their shortcomings through designing specialised schemes, cataloguing and subject indexing methods. Classification and cataloguing system have to work in tandem for better organisation and retrieval of information stored in our retrieval systems.

Some dedicated individuals developed a number of special classification schemes with their enthusiasm in this regards. But after their retirement from the profession, there were little revisions in their schemes. On the other hand, some special classification schemes, supported by the professional organisations, are well maintained and revised frequently. The WIPO (World Intellectual Property Organization)-sponsored Patent Classification System and ACM (Association of Computing Machineries)-sponsored Computing Classification Scheme are such examples, which bring out revisions at regular intervals to deal with growth of subjects. Some special classification schemes are also available in electronic formats, like on CD-ROM and Internet. This effort makes these classification schemes more accessible worldwide. Some depth classification and special classification schemes have also been introduced in organisation of electronic information resources, specially the Internet resources, to produce more relevant retrievals from the search services. Application of principles of classification in organizing Internet resources is now an emerging area of research and researchers in library and information science as well as the researchers in computer sciences are involved in this area. Compared to 1960s and 1970s, the development of new depth classification schemes and special classification schemes has lost its focus in organisations like, DRTC and CRG, due to some reasons or others.

Self Check Exercise

9) List the major trends in library classification studies & research.

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

.....

.....

.....

.....

.....

3.11 SUMMARY

Beginning with the DDC (1876) the modern classification systems are a little more than 125 years old. Since then, the DDC has inspired the development of many general classification systems for organisation of knowledge in libraries. Of these, about half a dozen are still alive, while a few others are internationally insignificant or dead. Of this family, the DDC is still the most visible and popular one. Despite its many weaknesses it is getting stronger and stronger. For many, it is the classification. It is being used by bout 200,000 libraries in 135 countries. At present, it is in its 22nd edition (2003) available in print and electronic formats. It is followed by a sister but rival system UDC adapted from the DDC in early 20th century. Formerly, a property of the FID, from January 1992, it has been owned by a multinational UDC Consortium (UDCC) running mostly on commercial lines. It is a first faceted classification, of and was designed for classification micro literature such as reports, patents, and journal articles, and for use in

bibliographies and catalogues. Available wholly or partially in about 20 languages, the 4th edition is in English published by the British Standards Institution. The International Medium Edition IME 1985-1988 and its revision (1993) which has 61,000 entries is currently most popular one. Standard edition with 61,000 entries is trilingual (English, French and German), version is available only in electronic form and his mother of all lingual and national editions. Pocket edition (1999) with only 4000 broader divisions is the latest one is available in machine readable form. Next in popularity is the Library of Congress Classification whose outline first published in 1905, is now in its 6th edition (1990). Being purely enumerative and very detailed classification with 21 main classes, it is huge in size having above 11000 pages in all contained in more than 48 volumes. It is a perfect example of an enumerative classification. It is popular in big research and university libraries in USA and abroad. Besides, the LC numbers are available on MARC record and CIP data. Being associated with a great library, it has an assured future, notwithstanding its many inherent structural faults. It is also available on a CD-ROM titled *Classification plus*. Published in USMARC format for classification data Classification plus contains schedules as well as LC Subject headings. Ranganthan's Colon Classification, unofficially projected as India's national schemes of classification, was first published in 1933. It brought a great breakthrough in library classification. It is now being replaced by the BC-2 as a model of allocated classification. Its sound postulates and principles, on which the CC is firmly based, have become part of the theory of classification in general. It is now in its 7th edition (1987). The new edition incorporates many far-reaching changes especially in postulating a large number of new basic subjects, and different varieties of common isolates. Some changes have been made in notation also. It is a less used classification. Its contribution lies in its methods and sound theoretical base which has revolutionised the theory and practice of classification. The Bibliographic Classification (BC), first designed by H.E. Bliss in 1940-1953 is now in its second edition popularly known as BC-2 edited by J. Mills. Though yet incomplete it is the most modern, practical and scientific classification available today. It is based on Ranganathan's theories of facet analysis and reflects the modern structure of knowledge. Its early success is striking. About 90 libraries in UK and elsewhere are using it. Its use is likely to grow. Its class 'J – Education' is considered the best classification of Education. Another rather colossal classification is the Russian BBK designed during 1961-1968. The full edition consists of 25 volumes but many other abridgments are available. It is a semi-faceted classification with many auxiliary tables modeled on the UDC. It was designed to classify knowledge from the Marxist-Leninist viewpoint. It is used in many libraries of the erstwhile USSR including some countries of the former Warsaw Pact, and also in Vietnam and Korea. With the decline of communism, its future is rather uncertain. Broad System of Ordering (BSO) is a system sponsored by UNSESCO for UNISIST programme and designed by FID/BSO panes headed by Eric J Coates. Consisting of broad 6800 divisions, it is not a classification scheme in the ordinary sense; nor was it designed to arrange books in libraries. It was designed as a roof classification or switching language, which means a link or conversion language or an intermediary to transfer bibliographical information from one indexing language to another, or from one information system to another. It is now in it 4th edition (1978), and available only in 3 floppies. No further interest is being taken in this unique system.

The history of classification shows a broad but unmistaken trend of progress from enumerative to faceted classification. Revised DDC, UDC and BC-2 are culmination of this trend. The CC has greatly influenced the designing of new and revision of the old classification systems. Special subject classification system are in vogue and surely faceted in approach. The most outstanding area of research now a days is to study the use of classification in online public access

catalogues (OPACs), electronic databases and Internet search engines. At present, no new general classification system is being designed while many scholars feel that it is high time to construct another general classification system to reflect the modern structure of knowledge and to organise electronic and networked information.

In this Unit you have also briefly learnt about the special classification systems. The general classification system deals with broader perspectives of universe of knowledge, whereas the special classification system deals with micro subjects or specialised forms of documents. You have learnt some features, scope, method of construction and examples of special classification systems in this Unit. The special classification schemes are used in some special libraries and information centres. A number of special classification schemes are also used in organisation of electronic information resources.

3.12 ANSWERS TO SELF CHECK EXERCISES

1) Some of the important salient features of DDC include:

   i) It is a classification by discipline; it divides the entire universe of knowledge into ten main classes. Each class is again subdivided by ten and so on using decimal notation.

   ii) This system uses pure and simple notation of Indo-Arabic numerals used decimally. Decimal notation is useful in depicting hierarchy of subjects and to accommodate always newly emerging topics.

   iii) Its efficient Relative Index which brings together topics scattered by discipline in the Schedules.

2) The advantages of WebDewey are: (i) continuous updation, (ii) many access points, and (iii) display options.

3) The UDC's different editions are:

   i) The UDC Complete Edition (2005)

   ii) The UDC Medium Edition (1993)

   iii) The UDC Abridged Edition (2003)

   iv) The UDC Online.

4) LCC is purely enumerative i.e. the class numbers are available readymade and there is little provision of synthesis. It is highly detailed. It is based on the literary warrant available at the Library of Congress.

5) A/Z (Roman Capitals) (DGreek Delta) 0/9 Indo-Arabic Numerals a/z (i, l, o excluded) (Roman smalls) indicator digits with anteriorising value,* (asterisk), "(double inverted comma) ← (backward arrow). Ordinary indicator digits are () (parenthesis), & (ampersand) '(inverted comma). (dot): (colon); (semi colon), (comma)- (hyphen) = (equal to) →(forward arrow) + (plus).

6) i) The order of main classes is based upon the educational and scientific consensus.

   ii) Notation is simple and short comprising of Roman Capitals and Indo-Arabic Numerals.

   iii) The system uses retroactive notation which can be combined backwardly.

   iv) The system is based upon modern facet analysis theories and methods first given by S.R. Ranganathan.

7) BBK was designed to represent the Marxist Leninist viewpoint in library classification. Its future is uncertain in the wake of the collapse of the Soviet Union.

8) The structure of BBK in Indo-Arabic numerals as follows:

1	Interdisciplinary knowledge.
2/5	Science, Technology. Agriculture and Medical sciences
6	Social Sciences
7	Culture and education
8	Language and literatures.
86	Religion, Atheism
87/88	Psychology
9	Universal subjects

9) Some of the major trends in library classification are to:

i) base classification system on the modes of growth of subjects;
ii) standardise terminology;
iii) design general classification to serve simultaneously as special and general classification systems. Most of the general classification systems are officially available as in editions of varying details;
iv) discover more varieties of auxiliary subdivisions;
v) make the notation simple but more versatile; and
vi) find uses of classification in electronic databases, OPACs and Internet search engines.

3.13 KEYWORDS

Auxiliaries : *See* Common Isolates.

Classification Plus : The Library of Congress classification schedules on CD-ROM.

Classification Schedule : The core part a library classification system which lists subjects in a systematic way along with their notation and some notes if necessary given against each.

Common Isolates/ Auxiliaries : Known by different names in different systems, these are usually non-subject but recurring aspects of knowledge as embodied in documents. These are not required in knowledge classification. These usually denote aspects of presentation of knowledge or media and format of documents.

Division by Discipline : Division of knowledge into broader areas of study such as Natural Science, Social Science, Humanities, and so on. All the modern classification systems now follow the DDC in dividing knowledge by discipline. James Duff Brown's Subject Classification followed the reverse method of classification by topic, but failed

Enumerative Classification : A classification that systematically lists all kinds of subjects of past, present and fore seeable future. Class numbers are available ready-made in such a system. The Library of Congress Classification and Rider's International Classification are at present two best examples of enumerative pattern

Faceted Classification : A species of classification that breaks a subject into some already postulated broad

categories and facets to be combined in a predetermined order according to given rules to tailor a class number for a given specific subject. As contrasted from an enumerative classification, it provides no ready-made class numbers for compound and complex subjects. Modern trend is towards faceted and analytico-synthetic classifications.

General Classification System : A classification whose area of purview extends to the entire domain of the universe of knowledge, and for classifying documents in all media. In contrast are the special classification schemes limited by subject area or by kind of documents e.g. classification of social sciences or classification for maps.

Hospitality of Notation : An essential quality of any notational system enabling it to accommodate new subjects at their logical places without at all disturbing the place of already existing subjects.

General Classification System : A classification whose area of purview limited to the narrower domain of the universe of knowledge, or for classifying documents in a particular media or format. In contrast the general classification schemes are expanded to universe of knowledge.

Switching Language : A switching language is one that enables "indexing decisions recorded in one indexing language or classification system to be 'translated' into equivalent decisions in another" – (Eric Hunter). BSO is one example of such a switching or exchange language for information transfer. It is also known as roof classification.

WebDewey : DDC-22 accessible through the Internet.

3.14 REFERENCES AND FURTHER READING

Chan, Lois Mai (1994). *Cataloguing and classification: an introduction*. 2nd edition. New York: McGraw-Hill.

Foskett, A.C. (1996). *The subject approach to information*. 5th ed. London: Library Association Publishing.

Gakhar, A.P. (1982). *Librarian's guide to Broad System of Ordering (BSO)*. New Delhi : Metropolitan.

Hunter, Eric J. (2002). *Classification made simple*. 2nd edition. Aldershot: Ashgate.

Krishan Kumar (1988). *Theory of library classification*. 4th edition. New Delhi: Vikas.

Marcella, Rita and Newton, Robert. (1994). *A new manual of classification* Aldershot: Gower.

Mcllwaine, I.C. (2000). *The Universal Decimal Classification: a guide to its use*. London: BSI Business Information.

Ranganathan, S.R. (1967). *Prolegomena to library classification*. 3rd edition, Assisted by M.A. Gopinath. Bombay: Asia Publishers.

Rowley, Jennifer and Farrow, John (2000). *Organizing knowledge*. 3rd edition. Aldershot: Gower.

Satija, M.P. (1989). *Colon Classification, 7th edition: a practical introduction*. New Delhi: Ess Ess Publishers.

Satija, M.P. (2001). *Exercises in the 19th edition of Dewey Decimal Classification*. New Delhi: Concept Publishing.

Sethi, A.R. (1985). BC through the prism of CC. *ILA Bulletin,* 21(3-4), 81-89.

Sukiasyan, E.R. (1988). Classification practice in the USSR: current status and development trends. *International Classification,* 15(2), 69-72.

Thomas, Alan R. (1997). Bibliographic Classification. *Cataloguing and Classification Quarterly,* 25(1), 51-104.

Universal Decimal Classification: *Abridged Edition*. (2003). London: BSI Business Information.

Universal Decimal Classification: *Complete Edition*. (2005). Volume 1: Systematic Tables, Volume 2: Alphabetical Index. London: BSI Business Information.

Websites

DDC	http://www.oclc.org/fp
BC-2	http://www.sid.cam.ac.uk/bca/bcahome.htm
LCC	http://clarssweb.loc.gov
UDC Consortium	http://www.udcc.org
UDC Online	http://www.udconline.net
UDC Pocket edition	http://www.bsi.org.uk.disc
UDC-Forum.html	http://www.jiscmal.ac.uk.lists/UDC-Forum.html
Email ID of CC	srels@vsnl.com
E-mail ID of BSO	eric@ejhunter.freeserve.co.uk

UNIT 4 INDEXING SYSTEMS AND TECHNIQUES

Structure

4.0 Objectives

4.1 Introduction

4.2 Indexing Principles and Process

- 4.2.1 Purpose of Indexing
- 4.2.2 Problems in Indexing
- 4.2.3 Indexing Process
- 4.2.4 Indexing Language
- 4.2.5 Theory of Indexing
- 4.2.6 Indexing Criteria
- 4.2.7 Indexing Policy: Exhaustivity and Specificity
- 4.2.8 Quality Control in Indexing

4.3 Pre-Coordinate Indexing Systems

- 4.3.1 Cutter's Rules for Dictionary Catalogue
- 4.3.2 Kaiser's Systematic Indexing
- 4.3.3 Chain Indexing
- 4.3.4 Relational Indexing
- 4.3.5 Coates's Subject Indexing
- 4.3.6 PRECIS
- 4.3.7 COMPASS
- 4.3.8 POPSI

4.4 Post-Coordinate Indexing Systems

- 4.4.1 Uniterm
- 4.4.2 Optical Coincidence Card / Peek-a-boo
- 4.4.3 Edge-Notched Card
- 4.4.4 Post-Coordinate Searching Devices

4.5 Automatic Indexing

- 4.5.1 Manual Indexing vs. Computerised Indexing
- 4.5.2 Methods of Computerised Indexing
- 4.5.3 File Organisation
- 4.5.4 Indexing Systems using AI Techniques
- 4.5.5 User Interface Design

4.6 Non-Conventional Indexing : Citation Indexing

4.7 Web Indexing

- 4.7.1 Meaning and Scope
- 4.7.2 Operational Aspects of the Web
- 4.7.3 Pre-requisites for Web Indexing
- 4.7.4 Search Engines
- 4.7.5 Semantic Web

4.8 Summary

4.9 Answers to Self Check Exercises

4.10 Keywords

4.11 References and Further Reading

ignou THE PEOPLE'S UNIVERSITY

4.0 OBJECTIVES

In the previous Units you have seen the importance of indexing language in information storage and retrieval. You have also learnt the characteristics and types of indexing languages. In this Unit we will be discussing the various indexing systems – their features and techniques for generating index entries for effective retrieval. The concept of Web indexing is also discussed.

After reading this Unit, you will be able to:

- understand principles, processes and problems of subject indexing;
- critically examine the significant contributions of various experts in subject indexing;
- learn to use the different indexing techniques according to requirements;
- explore the strengths and weaknesses of different indexing techniques;
- differentiate computerised indexing with manual indexing;
- understand different methods of computerized indexing;
- recognise different tools and techniques associated with the Artificial Intelligence-based subject indexing systems;
- appreciate the development of a non-conventional indexing technique – citation indexing and its use;
- understand the meaning and scope of the Web indexing;
- identify the features of different search tools used in finding Internet resources; and
- trace the development of technologies to make computers understand the semantics underlying Web documents.

4.1 INTRODUCTION

An index is a guide to the items contained in or concepts derived from a collection. Item denotes any book, article, report, abstract review etc. (text book, part of a collection, passage in a book, an article in a journal etc.). The word *index* has its origin in Latin and means: 'to point out, to guide, to direct, to locate'. An index indicates or refers to the location of an object or idea. The definition according to the British standards (BS 3700: 1964) is "a systematic guide to the text of any reading matter or to the contents of other collected documentary material, comprising a series of entries, with headings arranged in alphabetical or other chosen order and with references to show where each item indexed is located". An index is, thus, a working tool designed to help the user to find his way out mass of documented information in a given subject field, or document store. It gives subject access to documents irrespective their physical forms like books, periodical articles, newspapers, AV documents, and computer-readable records including Web resources.

Early indexes were limited to personal names or occurrences of words in the text indexed, rather than topical (subject concept) indexes. Topical indexes are found the beginning of the 18th century. In the nineteenth century, subject access to books was by means of a classification. Books were arranged by subject and their surrogates were correspondingly arranged in a classified catalogue. Only in the late 19th century, subject indexing became widespread and more systematic. Preparation of back-of-the-book index, historically, may be regarded as the father of all indexing techniques. Indexing techniques actually originated from these indexes. It was of two types: *Specific index*, which shows broad topic on the form of one-idea-one-entry, i.e. specific context of a specific idea; and *Relative*

index, which shows various aspects of an idea and its relationship with other ideas. Specific index cannot show this, it only shows broad topic on the form of one-idea-one-entry, i.e. specific context of a specific idea. The readymade lists of subject headings like Sears List and LCSH fall far short of actual requirement for depth indexing of micro documents in the sense that the terms are found to be too broad in the context of users' areas of interest and of the thought content of present day micro document.

4.2 INDEXING PRINCIPLES AND PROCESS

4.2.1 Purpose of Indexing

Indexing is regarded as the process of describing and identifying documents in terms of their subject contents. Here, The concepts are extracted from documents by the process of analysis, and then transcribed into the elements of the indexing systems, such as thesauri, classification schemes, etc.

In indexing decisions, concepts are recorded as data elements organised into easily accessible forms for retrieval. These records can appear in various forms, e.g. back-of-the-book indexes, indexes to catalogues and bibliographies, machine files, etc. The process of indexing has a close resemblance with the search process. Indexing procedures can be used, on one hand, for organising concepts into tools for information retrieval, and also, by analogy, for analysing and organising enquiries into concepts represented as descriptors or combinations of descriptors, classification symbols, etc. The main purposes of prescribing standard rules and procedures for subject indexing may be stated as follows:

- To prescribe a standard methodology to subject cataloguers and indexers for constructing subject headings.
- To be consistent in the choice and rendering of subject entries, using standard vocabulary and according to given rules and procedures.
- To be helpful to users in accessing any desired document(s) from the catalogue or index through different means of such approach.
- To decide on the optimum number of subject entries, and thus economise the bulk and cost of cataloguing indexing.

4.2.2 Problems in Indexing

A number of problems and issues are associated with indexing which are enumerated below:

a) Complexities in the subjects of documents—usually multi-word concept;

b) Multidimensional users need for information;

c) Choice of terms from several synonyms;

d) Choice of word forms (Singular / Plural form);

e) Distinguishing homographs;

f) Identifying term relationships — Syntactic and Semantic;

g) Depth of indexing (exhaustivity);

h) Levels of generality and specificity for representation of concepts (specificity);

i) Ensuring consistency in indexing between several indexers (inter-indexer consistency), and by the same indexer at different times (intra-indexer consistency);

j) Ensuring that indexing is done not merely on the basis of a document's intrinsic subject content but also according to the type of users who may be benefited from it and the types of requests for which the document is likely to be regarded as useful;

k) The kind of vocabulary to be used, and syntactical and other rules necessary for representing complex subjects; and

l) Problem of how to use the 'index assignment data'.

It is necessary for each information system to define for itself an indexing policy, which spell out the level of exhaustivity to be adopted, a vocabulary that will ensure the required degree of specificity-rules, procedures and controls that will ensure consistency in indexing, and methods by which users may interact with the information system, so that indexing may, as far as possible, be related to and be influenced by user needs and search queries. The exhaustivity and specificity are management decisions. Since document retrieval is based on the logical matching of document index terms and the terms of a query, the operation of indexing is absolutely crucial. If documents are incompletely or inaccurately indexed, two kinds of retrieval errors occur viz. irrelevant documents retrieval and relevant documents non-retrieval.

When indexing, it is necessary to understand, at least in general terms, what the document is about (aboutness). The subject content of a document comprises a number of concepts or ideas. For e.g. an article on lubricants for cold rolling of aluminium alloys will contain information on lubricants, cold rolling, aluminium alloys etc. The indexer selects these concepts, which are of potential value for the purpose of retrieval, i.e., those concepts on which according to him, information is likely to be sought for by the users. It is the choice of concepts or the inner ability to recognise what a document is about is in the very heart of the indexing procedure. However, it is the identification of concepts that contributes to inconsistencies in indexing.

The problem of vocabulary deals the rules for deciding which terms are admissible for membership in the vocabulary. There is also a problem of how to determine the goodness or effectiveness of any vocabulary. This implies that, the system rank each of the documents in the collection by the probability that it will satisfy given query of the user. Thus, the output documents relating to a search query are ranked according to their probability of satisfaction.

4.2.3 Indexing Process

Before indexing, the indexer should first take a look at the entire collection and make a series of decisions like,:

a) Does the collection contain any categories of material that should not be indexed?

b) Does the material require general, popular vocabulary in the index?

c) What is the nature of collection?

d) What is the characteristics of user population?

e) The physical environment in which the system will function; and

f) Display or physical appearance of the index.

Essentially, the processes of indexing consist of two stages: (i) establishing the concepts expressed in a document, i.e. the subject, and (ii) translating these concepts into the components of the indexing language.

a) **Establishing the concepts expressed in a document**

The process of establishing the subject of a document can itself be divided into three stages:

i) **Understanding the overall content of the document, the purpose of the author, etc**

Full comprehension about the content of the documents depends to a large extent on the form of the document. Two different cases can be distinguished, i.e. printed documents and non-printed documents. Full understanding of the printed documents depends upon an extensive reading of the text. However, this is not usually practicable, nor is it always necessary. The important parts of the text need to be considered carefully with particular attention to: title, abstract, introduction, the opening phrases of chapters and paragraphs, illustrations, tables, diagrams and their captions, the conclusion, words or groups of words which are underlined or printed in an unusual typeface. The author's intentions are usually stated in the introductory sections, while the final sections generally state how far these aims are achieved.

The indexer should scan all these elements during his study of the document. Indexing directly from the title is not recommended, and an abstract, if available should not be regarded as a satisfactory substitute for a reading of the text. Titles may be misleading; both titles and abstracts may be inadequate in many cases, neither is a reliable source of the kind of information required by an indexer.

A different situation is likely to arise in the case of non-printed documents, such as audio-visual, visual, sound media and electronic media.

ii) **Identification of concepts**

After examining the document, the indexer needs to follow a logical approach in selecting those concepts that best express its content. The selection of concepts can be related to a schema of categories recognised as important in the field covered by the document, e.g. phenomena, processes, properties operations, equipment etc. For example, when indexing works on 'Drug therapy', the indexer should check systematically for the presence or the absence of concepts relating to specific diseases, the name and type of drug, route of administration, results obtained and/or side effects, etc. Similarly, documents on the 'Synthesis of chemical compounds' should be searched for concepts indicating the manufacturing process, the operating conditions, and the products obtained, etc".

iii) **Selection of concepts**

The indexer does not necessarily need to retain, as indexing elements, all the concepts identified during the examination of the document. The choice of those concepts, which should be selected or rejected, depends on the purpose for which the indexing data will be used. Various kinds of purpose can be identified, ranging from the production of printed alphabetical indexes to the mechanized storage of data elements for subsequent retrieval. The kind of document being indexed may also affect the product. For example, indexing derived directly from the text of books, journal articles, etc. is likely to differ from that derived only from abstracts. However, the selection of concepts in indexing is governed by the *Indexing policy: exhaustivity and specificity* adopted by the given system (See Section 4.2.7 of this Unit).

b) **Translating the concepts into the indexing language**

In the next stage in subject indexing is to translate the selected concepts into the language of the indexing system. At this stage, an indexing can be looked from two different levels: document level, which is known as *Derivative indexing;*

and concept level, which is known as *Assignment indexing*. *Derivative indexing* is the indexing by *extraction.* Words or phrases actually occurring in a document can be selected or extracted directly from the document (keyword indexing, automatic indexing, etc.). Here, no attempt is made to use the indexing language, but to use only the words or phrases, which are manifested in the document. *Assignment indexing* (also known as '*concept Indexing')* involves the conceptual analysis of the contents of a document for selecting concepts expressed in it, assigning terms for those concepts from some form of controlled vocabulary according to given rules and procedures for displaying syntactic and semantic relationships (e.g. Chain Indexing, PRECIS, POPSI, Classification Schemes, etc.). Here, an indexing language is designed and it is used for both indexing and search process.

4.2.4 Indexing Language

An indexing language is an artificial language consisting of a set of terms and devices for handling the relationship between them for providing index description. It is also referred to as a *retrieval language.* An indexing language is 'artificial' in the sense that it may depend upon the vocabulary of natural language, though not always, but its syntax, semantics, word forms, etc. would be different from a natural language. Thus, an indexing language consists of elements that constitute its vocabulary (i.e. controlled vocabulary), rules for admissible expression (i.e. syntax) and semantics. More discussion on *indexing languages* can be seen in the Units 2 and 3 of this Course.

4.2.5 Theory of Indexing

The lack of an indexing theory to explain the indexing process is a major blind spot in information retrieval. Very little seems to have been written about the role and value of theory in indexing. Those who have written about it however, tend to agree that it serves a vital function. One important function of the theory of indexing is to establish agenda for research. Equally important, by identifying gaps it suggests what remains to be investigated. Theories also supply a rationale for, or an argument against, current practices in subject indexing. They can put things in perspective, or provide a new and different perspective.

The contributions made by K P Jones and R. Fugmann [Quinn, 1994] in indexing theory are worth mentioning. According to Jones, an indexing theory should consist of five levels, which are as follow:

a) *Concordance level:* It consists of references to all words in the original text arranged in alphabetical order.

b) *Information theoretic level:* This level calculates the likelihood of a word being chosen for indexing based on its frequency of occurrence within a text. For example, the more frequently a word appears, the less likely it is to be selected because the indexer reasons the document 'all about that'.

c) *Linguistic level:* This level of indexing theory attempts to explain how meaningful words are extracted from large units of text. Indexers regard opening paragraphs, chapters and/or sections, and opening and closing sentences of paragraphs are more likely to be a source of indexable Units, as are definitions.

d) *Textual level:* Beyond individual words or phrases lies the fourth level—the textual or skeletal framework. The author in his/her work presents ideas in an organized manner, which produces a skeletal structure clothed in text. The successful indexer needs to identify this skeleton by searching for clues on the surface.

e) *Inferential level:* An indexer is able to make inferences about the relationships

between words or phrases by observing the paragraph and sentence structure, and stripping the sentence of extraneous detail. This inference level makes it possible for the indexer to identify novel subject areas.

Indexing theory proposed by Robert Fugmann is based on five general axioms, which he claims have obvious validity and in need of no proof and they explain all currently known phenomena in information supply. These five axioms are:

a) *Axiom of definability:* Compiling information relevant to a topic can only be accomplished to the degree to which a topic can be defined.

b) *Axiom of order:* Any compilation of information relevant to a topic is an order creating process.

c) *Axiom of the sufficient degree of order:* The demands made on the degree of order increase as the size of a collection and frequency of searches increase.

d) *Axiom of predictability:* It says that the success of any directed search for relevant information hinges on how readily predictable or reconstructible are the modes of expression for concepts and statements in the search file. This axiom is based on the belief that the real purpose of vocabulary control devices is to enhance representational predictability.

e) *Axiom of fidelity:* It equates the success of any directed search for relevant information with the fidelity with which concepts and statements are expressed in the search file.

Like theories in other disciplines, these theories of indexing are developed provisionally, with the understanding that subsequent research will either support or refute them.

4.2.6 Indexing Criteria

It is possible, however, to minimise inconsistencies in indexing. Requiring that indexers systematically test the indexability of concepts by using a set of criteria can do this. It is obviously not possible to suggest criteria that would produce the same results when used by the same indexer at different times or by more than one indexer at the same time. The criteria at best enable greater agreement between indexers about concepts that should be indexed. Some of these criteria are given below in the form of a checklist of questions that indexers can ask themselves when faced with a document, to be indexed.

- To what extent the document is about a particular concept? Mere mention of any concept in the document does not make it indexable. If the concept was a reason for the document or if without the concept the document would either not exist or be significantly altered, then the concept is worth indexing.
- Is there enough information about the concept in the document? This is always a matter of judgment and indexers may disagree with one another about what constitutes 'enough information'. However, experience in indexing, in answering queries, and subject knowledge can go a long way in arriving at good decisions concerning this question.
- Another way of testing the indexability of a concept would be for the indexer to ask himself: would a user, searching for information on this concept, be happy if the document on hand is retrieved? Is there a likelihood of the concept figuring in search queries?

The answer to these questions would not only indicate the indexability of concepts but also the level of specificity at which concepts need to be indexed. To decide on the factors mentioned above, the indexer should have good judgment capacity, experience in answering search queries or reference service, good understanding of users and their information needs.

4.2.7 Indexing Policy: Exhaustivity and Specificity

Exhaustivity is a matter of an indexing policy and it is the measure of the extent to which all the distinct subjects are discussed in a particular document are recognized in indexing operation, and translated into the language of the system. Exhaustivity in indexing requires more number of index entries focusing different concepts (both primary and secondary) covered in the documents. The greater the number of concepts selected for indexing purpose, the more exhaustive is the indexing. If, in a given document, concepts A, B, C, D, E are selected for indexing then the indexing of the document is more exhaustive than if only concepts A< B< C are selected. When a relatively large number of concepts are indexed for each document, the policy followed is one of depth of indexing. Depth of indexing, in other words, allows for the recognition of concepts embodied not only in the main theme of the document but also in sub-themes of varying importance. Policy decision in respect of exhaustivity in indexing depends upon several factors like strength of collection, manpower available, economy and requirements of users.

In selecting a concept, the main criterion should always be its potential value as an element in expressing the subject content of the document. In making a choice of concepts, the indexer should constantly bear in mind the questions (as far as these can be known), which may be put to the information system. In effect, this criterion re-states the principal function of indexing. With this in mind, the indexer should:

- choice the concepts, which would be regarded as most, appropriate by a given community of users; and
- if necessary, modify both indexing tools and procedures as a result of feedback from enquiries.

Limit to the number of terms or descriptors, which can be assigned to a document should not be decided arbitrarily. This should be determined entirely by the amount of information contained in the document. Any arbitrary limit is likely to lead to loss of objectivity in the indexing, and to the distortion of information that would be of value for retrieval. If, for economic reasons, the number of terms is to be limited, the selection of concepts should be guided by the indexer's judgment concerning the relative importance of concepts in expressing the overall subject of the document.

In many cases the indexer needs to include, as part of the indexing data, concepts which are present only by implication, but which serve to set a given concept into an appropriate context.

Specificity is the degree of preciseness of the subject to express the thought content of the documents. It is the measure of the extent to which the indexing system permits the indexers to be precise when specifying the subject of the document. An indexing language is considered to be of high specificity if minute concepts are represented precisely by it. It is an intrinsic quality of the index language itself.

As a rule, concepts should be identified as specifically as possible. More general concepts may be selected in some circumstances, depending upon the purpose of the information retrieval system. In particular, the level of specificity may be affected by the weight attached to a concept by the author. If the indexer considers that an idea is not fully developed, or is referred to only casually by the author, indexing at a more general level may be justified.

Both *Exhaustivity* and *Specificity* are very closely related to recall and precision. A high level of exhaustivity increases recall and high level of specificity increases precision (See Section 5.4.1.1 under the Unit 5).

4.2.8 Quality Control in Indexing

The *quality* of indexing is defined in terms of its retrieval effectiveness—the ability to retrieve what is wanted and to avoid what is not. The quality of indexing depends on two factors: (i) the qualification of the indexer; and (ii) the quality of the indexing tools.

An indexing failure on the part of the indexer may take place at two stages of indexing process: establishing the concepts expressed in a document, and their translation. Failure in establishing concepts expressed in a document could be of two types:

a) Failure to identify a topic that is of potential interest to the target user group; and

b) Misinterpretation of the content of the document, leading to the selection of inappropriate term(s).

Translation failures may be of three types:

a) Failure to use the most specific term(s) to represent the subject of the document;

b) Use of inappropriate term(s) for the subject of a document because of the lack of subject knowledge or due to lack of seriousness on the part of the indexer; and

c) Omission of important term(s).

For a given information system, the indexing data assigned to a given document should be consistently the same regardless of the individual indexer. *Consistency* is a measure that relates to the work of two or more indexers. It should, remain relatively stable throughout the life of a particular indexing system. Consistency is particularly important if information is to be exchanged between agencies in a documentary network. An important factor in reaching the level of consistency is complete impartiality by the indexes. Almost inevitable, some elements of subjective judgment will affect indexing performance and these needs to be minimized as far as possible. Consistency is more difficult to achieve with a large indexing team, or with teams of indexer working in different location (as in a decentralized system). In this situation, a centralized check stage may be helpful.

The indexer should preferably be a specialist in the field for which the document is indexed. He should understand the term of the documents as well as the rules and procedures of the specific indexing system.

Quality control would be achieved more effectively if the indexers have contact with users. They could then, for example, determine whether certain descriptors may produce false combinations, and also create noise at the output stage.

Indexing quality is also dependent upon certain properties of the indexing method or procedure. It is essential that an index should be able to accommodate new terminology, and also new needs of users—that is, it must allow frequent updating.

Indexing quality can be tested by analysis of retrieval results, e.g. by calculating recall and precision ratios.

Self Check Exercises

1) Discuss the processes of subject indexing.

2) What do you mean by 'Exhaustivity' and 'Specificity' in indexing?

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

.....

.....

.....

.....

.....

.....

4.3 PRE-COORDINATE INDEXING SYSTEMS

4.3.1 Cutter's Rules for Dictionary Catalogue

It was Charles Ammi Cutter who first gave a generalised set of rules for subject indexing in his *Rules for a Dictionary Catalogue (RDC)* published in 1876. Cutter never used the term 'indexing'; rather he used the term 'cataloguing'.

Cutter identified two problems in organizing library materials in respect of developing subject headings: (a) problem relating to specific subject heading, and (b) order of terms in compound subject heading. Representation of the specific subject of a document by a single word or phrase appeared to be difficult in most cases. So, the complete analysis of the subject of the document is made and this situation paved the way for subject indexing from subject cataloguing. In order to solve the problems of constructing subject headings, Cutter, in his *RDC*, provided rules for specific as well as compound subject headings.

Specific Subject Heading

Cutter did not define the specific subject heading, but the idea was evident in Rule 161 of RDC. In regard to specific subject heading, Cutter advocated, *"Enter a work under its subject heading, not under the heading of a class which includes that subject"[Rule 161].* According to this rule, 'Roses' should be entered under the 'Roses' but not under the general category 'Flower'. However, Cutter gave the following example as a specific subject heading:

Subject of the Work	***Specific Subject Heading***
Movement of fluids in plant	Botany, Physiological

Here the work is entered under the heading of the class, which includes that subject, i.e. under broader class. Thus we see that in his own example Cutter could not maintain uniformity. Such inconsistency may be attributed to the fact that:

a) Cutter envisaged a set of 'stock subjects' under which every book had to be accommodated; and

b) Cutter could not visualize the subject which had no names or not nameable.

This suggests that though Cutter provided new idea for subject cataloguing but his examples were not commensurate with his ideas. There is a great difference between today's 'specific subject' and Cutter's 'stock subjects'. We should not forget the period of Cutter; the development of different subjects and multiple relationships between subjects which basically took place after the Second World

War. It was rather impossible to visualise and compile a list of 'stock subjects' to represent the subject of each and every document that is produced today.

Compound Subject Heading

Rule 175 of *RDC* states: *"Enter a compound subject name by its first word inverting the phrases only when some other word is decidedly more significant or is often used alone with the same meaning as the whole name".*

According to this rule, order of the component terms in compound subject heading should be the one that is decidedly more significant. But Cutter could not prescribe how one will come forward to decide which one is more significant. The question of significance varies from user to user. For example, a document on 'Social psychology' may be approached both by Sociologists and Psychologists. If this rule is followed, there would be two subject headings for the same document: (1) Psychology, Social; and (2) Social Psychology. Thus, we see that the decision in respect of 'significant' is left to the judgment of individual indexer, which is subjective one.

Cutter also provided some specific rules and guidelines in respect of the followings:

a) **Person vs. Country**

Entry will be under person in case of single/personal biography. Entry will be under country in case of history, event, etc.

b) **Country vs. Event**

Entry under event if it is proper noun. Entry under country if it is common noun e.g.

Freedom fighters in India: India, Freedom Fighters

c) **Subject vs. Country**

In scientific subjects, entry will be under the subject qualified by place:

Geology, India.

In areas, such as History, Government, Commerce, entry will be under place qualified by subject:

India—British Period.

For Humanities, Literature, Arts, etc., adjectival form is to be used:

Indian Painting.

d) **Between overlapping subjects,** entry will be according to the importance of the subject;

e) **Choice between different names**

- *Language*—If there are two languages out of which one is English, entry will be under English;
- *Synonyms*—Entry will be under one word with reference to the others;
- *Antonyms*—Entry will be under one word with reference to others;

f) **Compound Subject Heading:** The following examples show how Cutter tried to make a solution regarding compound subject heading. Each of these examples may have form and period subdivisions.

- A Noun preceded by an adjective, e.g. ***Political Economy, Ancient History;***
- A Noun preceded by another noun used as an adjective, e.g. ***Death Penalty, Flower Fertilization;***

- A Noun connected to another noun with a preposition—the direct form will be used, e.g. ***Patient with heart disease, Fertilization of flowers, Penalty of death;***
- Phrases or Sentence — Direct forms will be used, e.g. ***Medicine as Profession.***

g) **System of References** — Cutter advocated the following rules:

Rule 187: Make references from general subjects to their various subordinate subjects, and also to coordinate and illustrative subjects.

Rule 188: Make references occasionally from specific to general subject.

Cutter's system of references was from general to specific generally and specific to general occasionally. For the purpose of selecting subordinates for referencing, Cutter himself has prescribed the use of Dewey Decimal Classification and Expansive Classification.

Though Cutter tried to solve the problems of subject cataloguing of documents as it was existed, but he was not successful totally as he could not accommodate all the subjects in a single list and to make a uniform standard rules about the order of precedence of multi-worded subjects expressively.

4.3.2 Kaiser's Systematic Indexing

Julius Otto Kaiser systematised alphabetical subject heading practice by developing the principles behind Cutter's rules so as to form consistent grammar logic. Kaiser was the first person who applied the idea of Cutter in indexing micro documents in the library of Tariff Commission as its librarian. He started from the point where Cutter left. J. Kaiser in his "Systematic Indexing", published in 1911, pointed out that compound subjects might be analysed by determining the relative significance of the different component terms of compound subject through classificatory approach. He categorized the component terms into two fundamental categories: (1) Concrete and (2) Process. According to Kaiser,

Concrete refers to

- Things, place and abstract terms, not signifying any action or process; e.g. gold, India, Physics, etc.

Process refers to

- Mode of treatment of the subject by the author; e.g. *Evaluation* of IR system, *Critical analysis* of a drama.
- An action or process described in the document; e.g. *Indexing* of web documents.
- An adjective related to the concrete as component of the subject; e.g. Strength of metal.

Kaiser suggested a rule that a *'process'* should follow *'concrete'*. According to Kaiser, the *'concrete'* and *'process'* for the subject of a document on 'Heat treatment of metals' would be 'metals' and 'heat treatment' respectively and its subject heading would be: METALS—Heat Treatment.

Kaiser also envisaged that the 'concrete' is the main term, the 'process' gives the action state of the 'concrete', and the country supplies the locality where the action takes place. In the case of no 'concrete' and no 'place', the 'process' remains. But it is preferable to complete the subject heading with 'concrete' terms in all cases. As a result, we find the following unnatural subject headings:

Documents on	*Subject Headings*
Painting	PAINTS—Application
Painting of cars	CARS—Painting
Wage	LABOUR—Price
Strike	LABOUR—withdrawal

From the above examples it appears that Kaiser could not maintain consistency in the formulation of subject headings. He gave the idea of conceptual categorization but could not apply it properly in all cases. Beside this, Kaiser laid a rule that if a subject deals with place, double entry (*'Concrete—Place—Process'* and *'Place—Concrete—Process'*) is to be made. For example, index entries for 'Production of jute in West Bengal' would be: ***JUTE—West Bengal—Production; WEST BENGAL—Jute—Production.***

To help an enquirer in getting to the related topics, Kaiser recommended an elaborate system of references. He recommended that every 'concrete' term should be equipped with cross-references to both hierarchically superiors and inferiors. However, he did not recommend cross-references from 'process' terms.

In short, Kaiser's achievements regarding subject indexing are:

- *categorization* of composite terms through classificatory approach for the first time;
- a *general rule of order of precedence*, i.e. the 'process' term should follow the 'concrete' term;
- *definition* of those terms, of which 'process' is identified properly, that is, he gave the characteristics of 'process' by which it can be identified properly;
- *double entry system* for a subject dealing with place/locality; and
- elaborate *system of references*.

Though Kaiser brought much development in subject indexing, it had some drawbacks. These are:

a) There is no provision for entry under the 'process' term and as a result it fails to satisfy the users' approach by the 'process' term;

b) Double entry system recommended by him is uneconomical;

c) Kaiser's prescription of analysis of subject into *Concrete* and *Process* sometimes leads to unnatural headings; and

d) Kaiser left out the concept of 'time'.

Though Cutter and Kaiser could not solve many of the problems of subject indexing, still they could set the ball rolling to progress further to find out a logical solution to the problem of subject indexing.

4.3.3 Chain Indexing

Dr. S. R. Ranganathan developed a method of indexing, called *chain procedure* of subject indexing or simply Chain Indexing. It is a method of deriving alphabetical subject entries from the chain of successive subdivisions of subjects needed to be indexed leading from general to specific level. According to Ranganathan, chain indexing is a "procedure for deriving class index entry (i.e. subject index entry) which refers from a class to its class number in a more or less mechanical way." A note is also given with the above definition as:

"Chain procedure is used to derive class index entries in a Classified Catalogue, and specific subject entries, subject analytical, and 'see also' subject entries in a Dictionary Catalogue."

The term 'chain' refers to a modulated sequence of subclasses or isolates. Since the chain expressed the modulated sequence more effectively in a notational classification of subjects, this method takes the class number of the document concerned as the base for deriving subject headings not only for specific subject entry but also for subject reference entries. Therefore, the chain is nearly derived from a classification scheme, but not necessarily from a scheme. The method is intended to offer general as well as specific information to all information seekers by deriving subject headings from the chain of successive subdivisions that leads from the general to most specific level. While classifying a subject of a document we follow the same steps as those required for subject indexing in verbal plane (i.e. in natural language), the only difference is that in a classification scheme instead of representing the subject in verbal plane we use an artificial language to represent the subject in notational plane. The nature and structure of the classification scheme controls the structure of the subject headings drawn according to the chain procedure. The basic steps in chain indexing are enumerated below:

a) Classify the subject of a document by following a preferred classification scheme.

b) Represent the class number in the form of a chain in which each link consists of two parts: class number and its verbal translation in standard term or phrase used in the preferred classification scheme.

c) Determine different kinds of links: sought, unsought, false and missing links. Sought links denote the concepts (at any given stage of the chain) that the user is likely to access; unsought links are those that are not likely to be used as access points; false links are those that really do not represent any valid concept, mostly these are connecting symbols. Missing links represent those concepts that are not available in the preferred classification scheme, these are inserted by the indexer by means of verbal extension at the chain-with-gap corresponding to the missing isolate in the chain whenever there is such a need.

d) Derive specific subject heading for the specific subject entry from the last sought link and moving upwards by taking the necessary and sufficient sought links in a reverse rendering or backward rendering process. If the subject includes a space or a time isolate or a form, then break the chain into parts of the point(s) when the digit denoting a space or a time or a form. In such a situation specific subject heading is to be derived from last sought link of each part in reverse rendering process and during the reverse rendering process, the chain just before the part denoting time, space or form must be inverted, and the rest of the chain appended to it.

e) Derive subject reference heading for the subject reference from each of the upper sought links. This process will continue until all the terms of upper sought links are exhausted and indexed.

f) Prepare subject references or '*See also*' references from each subject reference heading to its specific subject heading when a subject heading starts from last sought link denoting space rendered, or a time or a form, prepare '*See*' references instead of '*See also*' references from subject headings to specific subject heading.

g) Prepare '*see*' references for each alternative and synonymous term used in the specific as well as subject reference headings.

h) Merge specific subject entries, subject references (i.e. '*See also*' references) and '*See*' references and arrange them in single alphabetical sequence.

The steps in chain indexing are demonstrated below with illustrative example:

1) *Subject Statement:* Researches in the promotion of student welfare programme through psychological counseling in Indian schools.

2) *Class no.:* 371.713072054 [according to DDC, 21st Edition]

3) *Representation of the class number in the form of a chain and determination of different kinds of links:*

300	Social Sciences
370	Education [SL]
371	Schools [SL] and their activities; Special education
371.7	Student welfare [SL]
371.71	Student health [SL] and related topics
371.713	Mental health services [SL]
371.7130	[FL]
371.71307	Study and Teaching [USL]
371.713072	Research [SL]
371.7130720	[FL]
371.71307205	Asia [USL]
371.713072054	India [SL]

4) *Preparation of specific subject heading, subject reference headings and cross references:*

Specific Subject Heading:

Research, Mental health services, Student welfare, Schools, India,

Subject Reference Headings:

Research, Mental health services, Student welfare, Schools

Mental health services, Student welfare, Schools

Student health, Student welfare, Schools

Student welfare, Schools

Schools, Education

Education

Cross References:

India, Research, Mental health services, Student welfare, Schools

See

Psychological counseling

Research, Mental health services, Student welfare, Schools, India

5) *Preparation of Index Entries*

Research, Mental health services, Student welfare, Schools, India
371.71307205

Bibliographical description and abstracts of the document are to be furnished under the specific subject heading

Research, Mental health services, Student welfare, Schools 371.713072

See also

Research, Mental health services, Student welfare, Schools, India

Mental health services, Student welfare, Schools 371.713

See also

Research, Mental health services, Student welfare, Schools, India

Student health, Student welfare, Schools 371.71

See also

Research, Mental health services, Student welfare, Schools, India

Student welfare, Schools 371.7

See also

Research, Mental health services, Student welfare, Schools, India

Schools, Education 371

See also

Research, Mental health services, Student welfare, Schools, India

Education 370

See also

Research, Mental health services, Student welfare, Schools, India

Psychological counseling *see* Mental health services

India, Research, Mental health services, Student welfare, Schools

See

Research, Mental health services, Student welfare, Schools 371.71302

6) *Alphabetisation*

Arrange the above entries according to single alphabetical order.

Chain indexing, a variety of pre-coordinate indexing, has some advantages like it is more or less a mechanical system and it is economic also. The greatest criticism against this method had been that only the last link provides the specific heading. Other entries are more general successively. Too much dependence upon classification scheme is another criticism. Ranganathan in a later paper clarified that index headings may be drawn both by forward rendering and backward rendering method. If we look deeply into the concept of Chain Indexing, it should be clear that Chain Indexing need not be dependent upon a particular classification scheme but it is based on classification principle and method.

4.3.4 Relational Indexing

J. E. L. Farradane devised a scheme of pre-coordinate indexing known as *Relational Indexing*. This indexing system was developed first in the early 1950s and has been modified since then. The latest change may be noted from Farradane's own paper in 1980. The basic principle of Farradane's Relational Indexing is to identify the relationship between each pair of terms of a given subject and to represent those relations by relational operators suggested by him and thus creating 'Analets'.

Relational operators are special symbols which link the isolates to show how they are related and each operator is denoted by a slash and a special symbol

having unique meaning. That is why the system propounded by Farradane is known as the *System of Relational Analysis*. He has formulated the basic idea of relational analysis by studying the development of learning process with particular reference to child psychology. Two or more isolates linked by *relational operators* as a subject statement constitute *Analet*.



Fig. 4.1: Analet

Two types of relationship has been identified and they are based on the qualities of experience of how children learn by developing power of discrimination in time and space and stages of discrimination in each of the following areas:

- The time sense involving three stage of discrimination towards complete association.
- Awareness of degree of distinctness (in space) involving three stages towards discrimination between terms.

In time, the first stage is 'non-time'—the co-occurrence of two ideas without reference of time; the second stage is 'temporary'—the co-occurrence of two ideas from time to time, but not permanently; and the third is 'fixed'—the permanent co-occurrence of two ideas. In space, the stages of discrimination are: the first is 'concurrent'— two concepts which are hard to distinguish i.e. parallel; the second is 'non-distinct'— two concepts which have much in common, e.g. cities and villages, science of religion; the third is 'distinct'—two concepts which can be completely distinguished.

Combination of these two mechanisms leads to nine categories of relations, which are shown through the following matrix:

Associative Mechanism (Increasing Association)

Discriminatory Mechanism (Increasing clarity of perception or recognition)

	Awareness (Non-time)	**Temporary Association**	**Fixed Association**
Concurrent Conceptualization	Concurrence / o	Self-activity / *	Association / ;
Non-distinct Conceptualization	Equivalence /=	Dimensional / +	Appurtenance / (
Distinct Conceptualization	Distinctness /)	Reaction / Action / ¾	Functional Dependence (Causation) /:

Fig. 4.2: Farradane's Relational Operators

Table 4.1: Farradane's Relational Operators with Illustrative Examples

Relational Categories and their operators	When to Use	Subject	Analet
1. Concurrence / 0	i) Mere co-existence of two terms like A in the presence of B; ii) Bibliographic form	Dictionary of Chemistry	Chemistry / 0 Dictionary
2. Self-activity / *	i) Intransitive verb; ii) Relation between the agent and activity	i) Migration of birds; ii) Broadcasting through satellite;	i) Bird / * Migration ii)Satellite /* Broadcasting
3. Association / ;	i) Unspecified association; ii) Relation of an agent, tool, thing/ application, discipline, indirect or calculated property, part or method of an action.	i) Application of computer in indexing; ii) Efficiency of calculator.	i) Indexing /; Computer ii) Calculator /; Efficiency
4. Equivalence / =	Equivalence of two things like synonyms, quasi-synonyms, something considered as something else, etc.	i) Vegetables like tomato; ii) Phosphate used as a fertilizer	i) Tomato / = Vegetables ii) Phosphate / = Fertilizer
5. Dimensional / +	Expresses position in space and time.	Petrochemicals plants at Haldia	Petrochemical Plants / + Haldia
6. Appurtenance / (	i) Whole-part relation; ii) Physical property belonging to a substance.	i) Floor of the building; ii) Density of acid;	i) Building / (Floor ii) Acid / (Density
7. Distinctness /)	Expresses the idea of an alternative or substitute	i) Photograph of tiger; ii) Gold and silver	i) Tiger /) Photograph ii) Gold /) Silver
8. Reaction/Action / –	Expresses action or effect of a thing or process on another thing or process	Purifying water	Water / – Purifying
9. Functional Dependence / :	i) Expresses the relation of one thing causing or producing something; ii) Product from a raw material or process.	i) Cyclone due to depression; ii) Extraction of minerals from ore	i) Depression / : Cyclone ii) Ore / : Minerals / – Extraction

Farradane's Relational Indexing offers no rule for filing order for the component terms in an index entry. If the reverse order of the components is required then the necessary operators are also required to be written in the reverse order. For example,

'Depression / : Cyclone' can also be written as *'Cyclone : / Depression'*.

Farradane has claimed that his system gives very good results in terms of recall and relevance, and that once the basic theory underlying the system is grasped, it can be applied very easily; it takes very little time to construct the *analet* once the subject of the document is determined. This claim is not borne out by other workers in the field. It was found that even after a fair amount of experience, the amount of time taken to index a document using this system was such as to make it highly uneconomic, while the results obtained were no better than with other systems. Farradane's Relational Indexing has been the subject of scholarly research, but was not implemented widely. Still, we can say that Farradane's marked improvement in the area of subject indexing was:

- analysis of relationship among terms;
- use of relational operators; and
- one to one relationship among analets.

4.3.5 Coates's Subject Indexing

Idea of E. J. Coates is not considered as original in nature. From the contributions of Cutter, Kaiser and Ranganathan, the concept of *Term Significance* was drawn. From the contribution of Farradane, the concept of *Term relationship* was drawn. Coates, in his contribution, has made a synthesis of above two concepts. It was advantageous for Coates to apply his idea on British Technology Index (now Current Technology Index) of which he had been the editor from its inception in 1963 until his retirement in 1976.

Term Significance

The most significant term in a compound subject heading is the one that is most readily available to the memory of the enquirer. From this, Coates has developed the idea of *Thing* and *Action* like Kaiser's Concrete and Process.

A *'Thing'* is whatever one can think, that is to say whatever can be thought as a static image. It is the image that comes straight into our mind, i.e. which we can visualize first. It includes not only the names of physical objects but systems and organisations of a mental kind, e.g. Democracy [system].

An *Action* refers to any thing in action or process denotes by term / word.

Example: Heat treatment of aluminium

ALUMINIUM [Action] / Heat treatment [Thing]

Among 'Thing' and 'Action', thing is more significant than Action. So the subject heading would be: Thing / Action. Coates also added the idea of M*aterial*. A *Material* is a state images produced by names of Things and names of Material. A Thing is made up of certain Material and the Material follows the Thing. A Thing has a boundary (i.e. shape) but the Material has none and for this reason the name of the Material is of low significance than the name of a Thing. On the other hand, the image of a Material is made of some static seeming properties—such as, colour, hardness, smoothness, etc. (i.e. the image of a material is produced due to its colour, hardness, smoothness, etc.) and is rated higher in the significance than the name of an *action.* Therefore, the significant order is *Thing—Material—Action* like Personality— Matter—Energy of Ranganathan. *Part* is a component of a thing. Hence, it depends on the thing to which they belong and, thus, giving us the order of significance: *Thing–Part–Material–Action.*

Term Relationship

The order of significance is fundamental in indexing. But it does not answer all questions of the order of the component terms. Apart from significance order, compound subjects may include two or more equally concrete things, or a complex subject may incorporate two or more phrases, which sometimes bring problems in indexing. These problems can be solved if the relationship between components of the compound term is considered. Establishing significance order for a subject heading requires a great deal of intellectual effort.

In order to decide the problems of above mentioned kind, it is necessary to consider:

a) How far relationships between components lead to modifications of the significance formula: *Thing – Material – Action*; and

b) How far the significance formula lead to modification of natural language order of the whole components.

Coates has provided a very valuable corollary of his ideas on significance order. Coates broke each phrase into terms with necessary preposition in between them which he called *Amplified phrase order*. Amplified phrase order is the order of component terms achieved by using the necessary prepositions in between them.

Examples:

Phrase / Compound Subject	***Amplified Phrase order***	***Subject Headings***
Conveyor belt	Belt of conveyor	Conveyor, Belt
Belt conveyor	Conveyor with belt	Conveyor, Belt

In a simple way, we can say that a prepositional phrase in the form of *Action–Preposition–Thing* is used while expressing the idea of a *Thing* being acted upon by an *Action*. Then we have to reverse the phrase omitting the preposition to generate a subject heading in accordance with the *Thing–Action* significance formula as shown in the example: ALUMINIUM / Heat treatment.

Coates has identified a number of relationships by means of a number of prepositions—such as, 'Of', 'For', 'Against', 'With' and 'By'. Coates distinguished 20 different kinds of relationships including *Thing – Action*, and tabulates these to show their relation to the corresponding prepositional phrase. Coates also suggested some rules in respect of order of precedence or order of significance.

Categories of Relationships

Categories of relationships proposed by Coates fall into following two broad divisions:

a) Relationships dealing with properties, actions, materials, viewpoints and parts in relation to one another and to Things [Categories 1—7].

b) Relationships dealing not with actions, parts, and materials as such but with Things differentiated by reference to parts, materials, principles of action, and causative factors [Categories 8—20].

Each category is illustrated below with examples:

Category 1: Action on Thing

Example: Planning of Villages

Subject Heading: VILLAGES [Thing], Planning [Action]

Significance formula: Agrees

Amplified phrase order: Reverses

Category 2: Action on Material

Example: Moulding of plastic

Subject Heading: PLASTIC [Material], Moulding [Action]

Significance formula: Agrees

Amplified phrase order: Reverses

Category 3: Action A on Action B

Example: Corruption in politics

Subject Heading: POLITICS [Action B], Corruption [Action A]

Significance formula: Nil (since no 'Thing' is present)

Amplified phrase order: Reverses

Category 4: Material of Thing

Example: Metal of container

Subject Heading: CONTAINERS [Thing], Metal [Material]

Significance formula: Agrees

Amplified phrase order: Reverses

Category 5: Part of Thing

Example: Skin of animals

Subject Heading: ANIMALS [Thing] Skin [Part]

Significance formula: Nil

Amplified phrase order: Reverses

Category 6: Property of Thing, Material and Action

Examples: i) Intelligence of man

ii) Hardness of iron

iii) Velocity of flow

Subject Headings: i) MAN [Thing], Intelligence [Property]

ii) IRON [Material], Hardness [Property]

iii) FLOW [Action], Velocity [Property]

Significance formula: Nil

Amplified phrase order: Reverses

Category 7: Partial viewpoint on Thing, Material or Action or Property.

Examples: i) Physics of metals (*Partial viewpoint on Thing*)

ii) Ethics of hunting (Partial viewpoint on Action)

Subject Headings: i) METAL [Thing], Physics [Viewpoint]

ii) HUNTING [Action], Ethics [Viewpoint]

Significance formula: Nil

Amplified phrase order: Reverses

Category 8: Things distinguished by citation of Principle of Action

Example: Bridges based on suspension

Subject Headings: BRIDGES [Thing], Suspension [Action]

Significance formula: Agrees

Amplified phrase order: Agrees

Category 9: Things distinguished by citation of Materials

Example: Bricks made of gold

Subject Heading: BRICKS [Thing,], Gold [Material]

Significance formula: Agrees

Amplified phrase order: Agrees

Category 10: Things distinguished by Part.

Example: Books with illustration

Subject Heading: BOOKS [Thing], Illustration [Part]

Significance formula: Agrees

Amplified phrase order: Agrees

Category 11: Thing distinguished by Material or Form of entry which it utilises.

Example: Engines operated by steam

Subject Heading: STEAM [Material] ENGINES [Thing]

Significance formula: Reverse

Amplified phrase order: Reverse

Category 12: Action A distinguished by citation of contributing or underlying Action B

Example: Welding by resistance

WELDING [Action A], Resistance [Action B]

Significance formula: Nil (Because both of them are Action)

Amplified phrase order: Agrees

Category 13: Thing A, serving or supplying or aiming at Thing B

Example: Transistors for radio

Subject Heading: RADIO [Thing B], Transistors [Thing A]

Significance formula: Nil (Because both of them are Thing)

Amplified phrase order: Reverse

Category 14: Thing A or Action A distinguished from homonyms by the fact that it serves Thing B or Action

Example: Accelerators for composting

Subject Heading: ACCELERATORS [Thing A], Composting [Thing B]

Significance formula: Nil (Because both of them are Thing)

Amplified phrase order: Agrees

Category 15: Thing serving or instrumental to Action (i.e. one Thing is used as instrument for action)

Example: Machines for washing

Subject Heading: WASHING [Action] MACHINES [Thing]

Significance formula: Reverse

Amplified phrase order: Reverse

Category 16: Thing A caused by Thing B

Example: Craters caused by meteorites

Subject Heading: METEORITES [Thing B], Craters [Thing A]

Significance formula: Nil

Amplified phrase order: Reverse

Category 17: Thing caused by Action

Example: Crystals caused by condensation

Subject Heading: CRYSTALS [Thing], Condensation [Action]

Significance formula: Agrees

Amplified phrase order: Agrees

Category 18: Action caused by Thing

Example: Reaction caused by enzymes

Subject Heading: ENZYMES [Thing], Reaction [Action]

Significance formula: Agrees

Amplified phrase order: Reverse

Category 19: Action A caused by Action B

Example: Rain caused by depression

Subject Heading: RAIN [Action A], Depression [Action B]

Significance formula: Nil

Amplified phrase order: Agrees

Category 20: Thing or Action at a type of location

Example: Plants found in Water

Subject Heading: WATER [Location], Plants [Thing]

Significance formula: Nil

4.3.6 PRECIS

Dereck Austin developed PRECIS, the PREserved Context Index System, in 1974 as an alternative procedure for deriving subject headings and generating index entries for British National Bibliography (BNB) which since 1952, was following Chain Indexing. Two most important factors played significant role in looking for an alternative method, ultimately resulted in the development of PRECIS: i) ideas of replacing chain indexing technique of BNB; and ii) the decision of the British Library to generate computer produced BNB with all the indexes. While developing PRECIS the following factors were taken into consideration.

a) The computer, not the indexer, should produce all index entries. The indexer's responsibility would be only to prepare the input strings and to give necessary instructions to the computer to generate indexes according to definite formats;

b) Each of the sought terms should find index entries and each entry should express the complete thought content/full context of the document unlike the chain indexing where only one entry is fully co-extensive with the subject and others are cross references describing only one aspect of the complete content of the document;

c) Each of the entries should be expressive;

d) The system should be based on a single set of logical rules to make it consistent; and

e) The system must have sufficient references for semantically related terms.

In the light of the experiences gained through the application of PRECIS in UK, USA, Canada, Australia, Germany and in counties as linguistically and culturally far apart, as Denmark, Italy, Poland, China, etc., Dr. Dereck Austin brought out its second edition in 1984.

Syntax and Semantics of PRECIS

PRECIS consist of two inter-related sets of working procedures: *Syntactical* and *Semantic*. Syntactical relationships in PRECIS are handled by means of a set of logical rules and a *schema of role operators* and *codes*. They refer to the organisation of terms in input strings and their manipulation to generate index entries.

Semantic, i.e. a *priori* relationship between indexing terms and their synonyms are regulated by machine-held thesaurus that serves as the sources of *see* and *see also* references in the index. A thesaurus is generated simultaneously with the preparation of input string. The PRECIS manual goes into great and technical details about the online construction of input record for establishing semantic network of terms.

Principles of PRECIS

The PRECIS is based on two principles:

a) *Principle of Context Dependency:* By name, it suggests that the PRECIS is based on the *Principle of Context Dependency.* Each of the terms in the input string sets the next term into its most obvious wider context (i.e. general to specific). In other words, the meaning of each term in the string depends upon the meaning of its preceding term and taken together, they all represent a single context.

b) *Principle of One-to One Relationship:* When terms are organised in context-dependent order, they form *a one-to one relation.* This simply means that each of the terms in the string is directly related to its next term.

Thus, these two characteristics (*context dependency*, leading to terms linked by *one-to-one relations*) are the basis on which the whole system works and they play a very important role in conveying the meaning of an index entry.

Entry Structure of PRECIS

In order to achieve the principle of context-dependency, *Two-Line-Three-Part* entry structure is followed in PRECIS and this can be demonstrated with the following diagram:

A string of five terms A-B-C-D-E has been represented in one-to-one related sequence. The term A or the term E can act as entry point. But, when C is considered as entry point, the different structure is necessary. A '*Two-Line-Three-Part'* entry structure expresses the relationship of the terms fully (Figure 4.3).

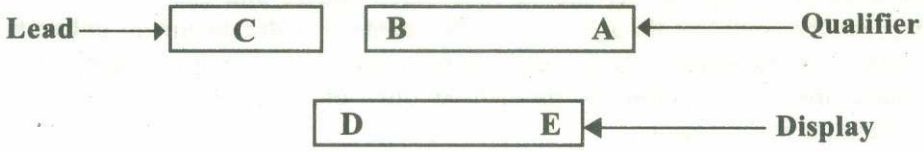


Fig. 4.3 : Entry structure

The first line consists of two units—the *lead* and *qualifier*—is called the *heading,* and the other line is called the *display. Lead* is occupied by approach term or filling term and is offered as the user's access point in the index. *Qualifier* position is occupied by the term(s) that sets the lead into its wider context (i.e. general to specific). Heading consists of terms set down in a narrower-to-wider context order. When *Lead* is occupied by the first term of the input string, the qualifier position is usually kept blank. *Display* position is occupied by those terms that rely upon the heading for their context. Sometimes, the display position becomes empty when lead is occupied by the last term of the input string.

Input String, Role Operators and Codes

It has already been stated earlier that the grammar of PRECIS is based on the codes and role operators, which form the syntactical backbone of PRECIS and regulate the writing of input string. *Input String* refers to a set of terms arranged according to the role operators which act as instructions to the computer. *Role Operators* refer to a set of notations, which specifies the grammatical role or function of the term which follows the operators and which regulates the order of the terms in an input string. The rules associated with role operators serve as computer instructions for generating index entries determines the format, typography and punctuation associated with each index entry.

There are two kinds of Role Operators: *Primary Operators* (earlier called Mainline Operators) and *Secondary Operators* (earlier called Interposed Operators). One of these operators has to be written as a prefix to each of the terms in an input string. *Primary Operators* control the sequence of terms in the input string and also determine the format of entries in the printed index. There are seven primary operators in the range 0 to 6 having built-in filing value for three sets of concepts: (0) is for the 'environment of the core concepts'; (1), (2) and (3) are for the 'core concepts'; and (4), (5) and (6) are for the 'extra-core concepts'. *Secondary Operators* (earlier called 'Interposed Operators') can be introduced into a string at any point to raise its level of exhaustivity, but these operators cannot be used to start a string. These operators are used for three sets of concepts: (f) and (g) are for 'coordinate concepts'; (p), (q) and (r) are for 'dependent elements'; and (s), (t) and (u) are for 'special classes of action'. Any of the secondary operators is always to be preceded by a primary operator to which it relates. The term which has to come in lead position is to be marked by (✓).

An important feature of the revised version of PRECIS is the provision of *Codes* for bringing expressiveness in the resulting index entries. Three types of codes are there – Primary, Secondary, and Typographic codes. There are two sets of primary codes: one set (*theme interlinks*) involves three codes used to link common or related themes in the subject statement, while the other (*term codes*) also involves three codes, used to denote the status of a term in the subject statement. There are also some secondary codes which are divided into two sets: the first set, *differences,* involves three kinds of differences—such as, 'preceding

differences', 'date as a difference' and 'parenthetical differences'. The second set involves *connectives,* which are used to connect any two or more consecutive terms in the input string in order to bring expressiveness in the resulting index entries. There are also some *typographic codes* that are used to bring the desired typographic form of a given term in the resulting index entries.

Schema of Role Operators

Primary Operators

Environment of core concepts	0	Location
Core concepts	1	Key System Thing when action not present. Thing towards which an action is directed, e.g. object of transitive action, performer of intransitive action
	2	Action; Effect of action
	3	Performer of transitive action (Agent, Instrument); Intake; Factor
Extra-core concepts	4	Viewpoint-as-form
	5	Selected Instance: study region, study example, sample population
	6	Form of document; Target user

Secondary Operations

Co-ordinate concepts	f	'Bound' co-ordinate concept
	g	Standard co-ordinate concept
Dependent elements	p	Part; Property
	q	Member of quasi-generic group
	r	Assembly
Special classes of action	s	Roll definer; Directional property
	t	Author-attributed action
	u	Two-way interaction

Codes used in PRECIS Strings

Primary codes

Theme Interlinks	$x 1st concept in coordinate theme
	$y 2nd/subsequent concept in co-ordinate theme
	$z Common concept
Term Codes	$a Common noun
	$c Proper name (class of-one)
	$d Place name

Secondary codes

Differences

Preceding differences 1st and 2nd characters:

(3 characters)

$0 Non-lead, space generating
$1 Non-lead, close-up
$2 Lead, space generating
$3 Lead, close-up

3 rd character = number in the range, 1 to 9 indicating level of difference

Date as a 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 in italic preceded by full point
$I Filing part in italic, no preceding punctuation

Steps in PRECIS

Major steps involved in indexing according to PRECIS include the following:

1) Analysing the thought content of the document concerned and identifying the component terms denoting key concepts.
2) Organizing the component terms into a subject statement based on the principle of context dependency.
3) Determining the role or status of each term in terms of role operators.
4) Assigning the role operators, which signify the syntactical role of each term.
5) Deciding which terms should be the access points and which terms would be in other positions in the index entries, and assigning further codes to achieve these results.
6) Adding further prepositions, auxiliaries or phrases, which would result in clarity and expressiveness in the resulting index entries.
7) Generation of index entries by the computer.
8) Generation of supporting reference entries (i.e. *'see'* and *'see also' references*) from semantically related terms taken from a machine-held thesaurus.

Formats of PRECIS Index

There are three kinds of format in PRECIS: *Standard Format, Inverted Format* and *Predicate Transformation.* Each format is related to one or more role operators (primary) and is used while dealing with the appropriate role operator(s).

Standard Format

Index entries in the *standard format* are generated with the primary operators (0), (1) and (2) through the process, known as *Shunting*, by the computer. Here, a string of terms is marshaled, like railway tracks, in the display position prior to generation of any index entry. As the shunting process begins, the computer will generate the first index entry by pushing the first term of the input string into the lead position. Pushing the second term into the lead position and thereby replacing the existing lead term into the qualifier position will generate the second index entry. The process will continue until the computer reaches the last term of the input string. As soon as any term goes to the lead position, it is printed in bold typeface.

Example: 'Training of labours in India

(0) ✓ India

(1) ✓ labours

(2) ✓ training

India

Labours. Training

Labours. India

Training

Training. Labours. India

Inverted Format

Index entries in the *inverted format* are generated whenever a term coded by an operator in the range from (4) to (6) or its dependent elements appear in the lead. The rule relating to the generation of index entries with this format is that—when any of the terms coded (4), or (5) or (6) or any of their dependent element operators appear in the lead, the whole input string is read from top to bottom and is written in the display. However, if the term appearing in the lead is last term of the input string, then it will be dropped, from the display.

Example: 'A bibliography of statistics for librarians'

(1) ✓ Statistics

(5) ✓ Librarians $01 for

(6) ✓ bibliographies

Statistics

– *For librarians – Bibliographies*

Librarians

Statistics – *For librarians – Bibliographies*

Bibliographies

Statistics – *For librarians*

Predicate Transformation

When an entry is generated under a term coded (3) that immediately follows a term coded either by (2) or (s) or(t)—each of which introduces an action of one kind or another—the *predicate transformation* takes place.

Example: 'In-service training of librarians by teachers

(1) ✓ librarians
(2) ✓ training $21 In-service $v by $w of
(3) ✓ teachers

Librarians
In-service training by teachers

Training. Librarians
In-service training by teachers

Teachers
In service training of librarians

The following example illustrates the input string for a given subject statement and the generation of PRECIS index entries using different operators:

<u>Subject Statement</u>: A report on policies of the Government of India of the applications of computers in manpower planning of universities.

<u>Input String</u>:

(1) ✓ universities

(p) ✓ manpower

(2) ✓ planning

(sub 2 ↑) (2) ✓ manpower planning $w of

(s) ✓ applications $v of $ in

(3) ✓ computers $w for

(s) ✓ policies $v of $w on

(sub 2 ↑) (3) ✓ Government of India

(1) ✓ India

(2) ✓ Government

Annotation

1) The term 'applications' coded (s) introduces the idea of tool or instrument, used by an unspecified agent, generally human. This role defining term explains the function of the 'computer' in relation to 'manpower planning'.

2) The term 'policies' refers to an attribute (a property) of the term cited next (i.e. Government of India) and is directed towards the 'applications of computers in manpower planning of universities'.

3) Connectives $v and $w have been used in the string to construct pre-coordinate phrases in the resulting index entries.

4) Upward reading substitutes "Manpower planning' and downward reading substitutes 'Government of India' are preceded by the operators (2) and (3) respectively in order to express the role of the phrase as a whole, since they will affect the position of the substitute in the resulting index entries. While generating index entries, downward reading substitute and upward reading substitute are always ignored when the string is read in upward and downward directions respectively.

5) When the term coded (s) goes to the lead, the term coded (3) will go to the qualifier and the concept directed will go to the display. The extent to which the terms coded (s) justify making as leads depends on the subject field and also on the local practice.

Index Entries:

Universities

Manpower. Planning. Applications of computers.

Policies of Government of India

Manpower. Universities

Planning. Applications of computers. Policies of Government of India

Planning. Manpower. Universities

Applications of computers. Policies of Government of India

Computers

Applications in manpower planning of universities.

Policies of Government of India

Policies. Government of India

On computers for applications in manpower planning of universities

India

Government. Policies on computers for applications in manpower planning of universities

Government. India

Policies on computers for applications in manpower planning of universities

Annotation

In the above examples, Lead and Qualifier are separated by a full stop and two letter spaces. The standard separator between two terms in the entry is full stop

and one space. However, for certain operators, other punctuation's like colon, space, dash, etc. are used. The display is written leaving 2-letter space from the left. For over-run of display, 4-letter space and for over-run of heading 6-letter spaces are left from the margin.

4.3.7 COMPASS

In 1990, it was decided to revise UKMARC and to replace PRECIS by a more simplified system of subject indexing. As a result Computer Aided Subject System (COMPASS) was introduced for BNB from 1991 using the same kind of basic principles of PRECIS and the PRECIS was dropped.

PRECIS was designed for the specific purpose of generating coextensive subject statement at each entry point in a form that was suitable for a printed bibliography. This was not necessarily the best format for online searching. It is worth recapitulating briefly some of the criticism of PRECIS in order to put the COMPASS into proper perspective:

i) The syntactic structure of PRECIS is complex and time consuming. Their complex systems of role operators served to provide the output string for printing, but were not otherwise utilized. Through there is no reason why they should not have been.

ii) PRECIS appears to be imprecise in some aspects; for example, in many instances it does not appear to make any difference whether a concept is coded (1) or (2), which suggests that the operators would not be of much help in searching a computer file, where they might be included.

iii) Place name has been treated in several ways as part of the subject string. Depending on the sense, place name is coded by the operators (0), (1), (5) and occasionally (3).

iv) The author information may be of value, if an individual or a corporate body is closely associated with a particular subject. Persons as subjects, for examples of biographies, also from part of the PRECIS subject string. As a result entries for an individual may be found in both the Author/Title file and the subject file. Common practice for many years has been to file such entries in the Author/Title file, taking this as a name file .If a record is being searched online, it is immaterial whereas in the records a piece of information occurs, so long as it is found if it is there.

v) PRECIS allows very long headings. For example,

Acquisition. Books. Stock. Libraries. Universities. United States Selection. Approach plans – *Reports*

In this example, everything before "selection" is the heading. Long heading like this are not likely to be shared by more than one index element and the main purpose of distinguishing headings from subheadings seems to be thwarted. Even when more than 100 index elements begin with "Acquisition", a PRECIS index display will repeat this term each time if the other component terms of the heading are different. By contrast, in a system in which the lead term alone always forms the heading, the lead term "Acquisition" could be displayed once for a number of the index elements.

vi) PRECIS index generation rules are quite complex. It is practically difficult for an indexer to keep nearly 200 rules in mind every moment.

vii) User of the PRECIS manual (be he a student, teacher, or a practicing indexer) is too often confronted by the fine distinctions and interpretations, which sometimes seem incomprehensible and, thus, leads to consistency.

Syntactical and Semantic aspects of Compass

Role operators used in COMPASS are similar to PRECIS role operators. In order to minimize the complexity of PRECIS role operators, primary operator's (1), (2) and (3) are used for COMPASS along with the secondary operators' (p), (q), and (r). The PRECIS primary operators' (0), (4), (5) and (6) are not used for COMPASS. Codes in respect of "connectives" –that is, $v and $w are used in the input string to construct pre-coordinate phrases and thus to disambiguate the resulting index entries. In a subject "Assessment of students by teachers", for example, the following index entry:

Students
Assessment. Teachers

may lead to represent another subject "Assessment of teacher by students". In such cases, use of "connectives" with the term (i.e. Assessment) dealing with the concept of actions in the input string will reveal the subject properly from the resulting index entry as displayed below:

Students
Assessment by teacher

Pre-coordination also has a role to play in an on-line information retrieval system. An on-line search is usually based on keywords linked by the standard Boolean Operators (AND, OR, NOT). These operators allow the user to identify records where a given set of terms co-occurred in designated fields, but they cannot show how the concepts are interrelated .The on-line search of UKMARC file with the search statement "Teacher AND Student AND Assessment " will retrieve the documents on two different subjects as stated above. This can be largely avoided if the user can call for the display of a pre-coordinate subject statement as a kind of filter, during search procedure.

Like PRECIS dates as a difference (coded with $d) are not used in all cases. Historical periods may be listed in the input string from the DDC schedule.

In Literature, for example, the periods subdivisions enumerated in the DDC are used so that the index entry and the class number coincide. The methods associates with the generation of COMPASS index entries are same as that of PRECIS index entries.

The COMPASS is based on two basic components:

1) Subject Authority File; and
2) Thesaurus containing the semantic network of related terms, which serve as a source for *'see'* and *'see also'* references.

1) **Subject Authority File**

Subject Authority File developed by using the Washington Library Network (WLN) software contains two types of records:

i) *Term Authority Records (TAR)*, which contain terms representing single concepts—topical (including forms) or geographic (as subject and also as forms). Here, term may be either unitary (single-word) or compound (multi-word) expression of simple subject concepts. As for examples,

Unitary term heading: *Drugs*
Compound term heading: *Drug abuse*

ii) *String Authority Records (SAR)*, which contain one type of string only (i.e. topical). Each string consists of combination of two or more terms expressing complex subject. As for example,

String heading: *Drug abuse. Prevention*

A subject string once assigned may be reused for indexing documents of the same subject. Terms taken from TAR are arranged alphabetically under the given term along with the DDC number. The subject index of BNB refers to a DDC number in the following manner:

Hinduism
Jainism *related to* Hinduism 200

The above index entry looks like a relative index, first devised by Melvil Dewey.

In the main part of BNB, the COMPASS string appears at the end of the entry for bibliographic record. The printed subject index of BNB drawn according to COMPASS appears to be much shorter and user-friendlier at the expense of the precision.

2) **Thesaurus**

The semantic aspects of COMPASS are governed by the PRECIS network of related terms (i.e. the RIN file), which served as the source of *"see"* and *"see also"* references in the printed subject index of BNB. Subject Authority File developed for COMPASS by using the software WLN is also based on the concept of open–ended vocabulary, which mean that a new term can be admitted into the indexing vocabulary as soon as it is encountered in documents. This file also incorporates entries from the RIN file as appropriate.

COMPASS and UKMARC

When a document is handled by the office responsible for descriptive cataloguing, its detail is recorded on a standard worksheet, organised by MARC fields. Not all the information in the MARC record is relevant to the subject of a document, but certain fields are related specifically to classification number and others are related to subject headings. The different fields of the worksheet containing subject information of BNB are as follows:

050 Library of Congress (=LC) classification number
082 Dewey Decimal Classification (=DDC) number
600 Personal name as subject
610 Corporate name as subject
650 Topical subject headings
651 Geographical subject headings
660 Subject topical descriptors
661 Subject geographical descriptors

Fields 690, 691 and 692 used earlier for PRECIS string, RIN and Subject Indicator Number (SIN) respectively were discontinued: and the fields 660 and 661 replaced, with the introduction of COMPASS. Proper names as subject are located in the UKMARC fields 600 and 610 and are found in the Author/Title index, marked "t" to distinguish them from the same name as author or title. With the introduction of COMPASS, BNB discontinued inclusion of Library of Congress Subject Headings (LCSH), which was constructed from 6XX fields in the MARC records. LCSH contained headings with subdivisions, and headings subdivided according to "pattern headings" — neither of which was in the printed version of LCSH.

COMPASS and DDC

The indexer who writes the COMPASS input string also assigns the appropriate DDC number in the 082 field. The initial step of subject analysis is done only once while preparing the COMPASS input string for a document and this input string is taken as the basis for all later decisions relating to the derivation of the

subject data for a given document, and their incorporation in the relevant fields in the worksheet meant for the BNB. The DDC numbers serve three different purposes:

a) DDC numbers organise entries for bibliographic records in the classified part of the BNB and thus ensure collection of entries.

b) DDC numbers are also used as a source of feature headings. Usually, up to three levels of headings from DDC number are given, but sometimes for four and even five levels of heading from the DDC number are given. There are no feature headings for more details subdivisions. The index string drawn according to the COMPASS appears at the end of the main body of the entry, in italics. This system of producing feature heading has been reported to be unsatisfactory from the user point of view. In this connection, it is to be pointed out here that prior to the introduction of COMPASS, feature headings were constructed by the computer from the terms in PRECIS string selected in input order.

c) DDC numbers are now linked directly to the bibliographic records rather than through the subject string. Prior to the introduction of COMPASS, the PRECIS strings were used to generate the DDC numbers and also the feature heading for the BNB classified sequence, so that users could find their way through the file without being unduly reliant on the class numbers — both these are incorporated in the SIN file. The COMPASS has not been used for this purpose at all.

COMPASS and BNB

As stated earlier, the COMPASS string appears at the end for bibliographic record of a document in the classified/main part of the BNB. The subject index of BNB refers to a class numbers in the following manner:

Library operations
Classification compared with indexing 025

In the classified part of the BNB, a number of entries or bibliographic records has been arranged under the class numbers 025. The above mentioned subject index directs the user to scan entries under the class number 025 in the subject/ main part of the BNB in order to find out the one which has at the end of the subject heading "Classification compared with indexing".

The printed subject index of BNB appears to be much shorter and more user-friendly than the earlier one. A C Foskett has observed that in the BNB of 1990 (using PRECIS index), there are ten "see also" references for a term "Antiquities" and these are followed by 54 subject index entries, each modified by terms either in the qualifier or in the display to give a coextensive subject statement, leading to 54 different classes. In the BNB of 1991 using COMPASS index, the heading "Antiquities" appears at 28 places in the classified sequence leading to 28 different class numbers. To help the user, the subject index gives the context in which a particular piece of notation represents a given term. This looks like a relative index first devised by Melvil Dewey. In effect, every entry in the index is a "see" reference, taking us from the concept in verbal plane to its notational plane—that is, heading used for arrangement of entries for bibliographic records in the classified or main part of the BNB.

Any system needs time for its testing and development. With the introduction of COMPASS, BNB stopped including LCSH headings until protests from customers

and cooperative partners which finally led to their reintroduction in 1995. At the end of 1996 the British Library's own COMPASS system of indexing was dropped from the BNB with little or no protest from anybody and LCSH was further adopted as its sole subject access system.

4.3.8 POPSI

In 1964, Ranganathan demonstrated a new line of thinking regarding verbal indexing based on facet analysis [Ranganathan, 1964]. Since then, continuous research in this new line of thinking was going on, and ultimately the researches of Dr G Bhattacharyya on this new line of thinking led to three distinct but interrelated contributions in the field of subject indexing in India, which together constitute what has been designated as the General Theory of Subject Indexing Language (GT-SIL). Dr Bhattacharyya first explained the fundamentals of subject indexing languages with an extensive theoretical background in the GT-SIL. Three interrelated contributions, which constitute the GT-SIL, are as follows:

1) A methodology for structural analysis of names-of-subject developed through logical abstraction of the structures of outstanding SILs—such as those of Cutter, Dewey, Kaiser and Ranganathan. This methodology was based on the Deep Structure of Subject of Indexing Languages (DS-SILs).
2) A methodology for designing a specific purpose-oriented SIL, known as POstulate-based Permuted Subject Indexing or POPSI. The POPSI was developed through logical interpretation of the postulates for structural analysis of the names-of-subject forming part of the GT-SIL.
3) A methodology for designing a vocabulary control device, known as classaurus—an elementary category-based faceted scheme of hierarchical classification in verbal plane having all the necessary attributes of a conventional alphabetical thesaurus. It was designed on the basis of the postulates governing the structural analysis of names-of-subject forming part of the GT-SIL.

Structures of Subject Propositions

For the purpose of designing a SIL, the following types of structure in a name-of-subject (subject–proposition) have been recognized by Dr. Bhattacharyya:

1) ***Semantic structure:*** It refers to the structure in the dimension of denotation or comprehension. This structure is intrinsic to subject-propositions and is based on "Genus- Species", "Whole-Parts", and "Interfacet" relationships.
2) ***Elementary structure:*** It is the structure in the dimension of elementary categories to which the different substantive constituents of subject-propositions belong. This structure is artificially postulated and is recognized on the basis of the semantic significance of the substantives.
3) ***Syntactic structure:*** It is the structure in the dimension of the horizontal sequence of the elementary constituents of subject-propositions meant to generate the intended pattern of grouping in their vertical sequence.

Deep Structure of SIL (DS-SILs)

According to the GT-SIL, the structure of a specific SIL may be deemed to be a surface structure of the deep structure of SIL. By logically abstracting the results of analyses of the surface structures of the different SILs, it has been possible to arrive at the deep structure of SIL. The DS-SILs may be presented schematically as shown in Figure 4.4.

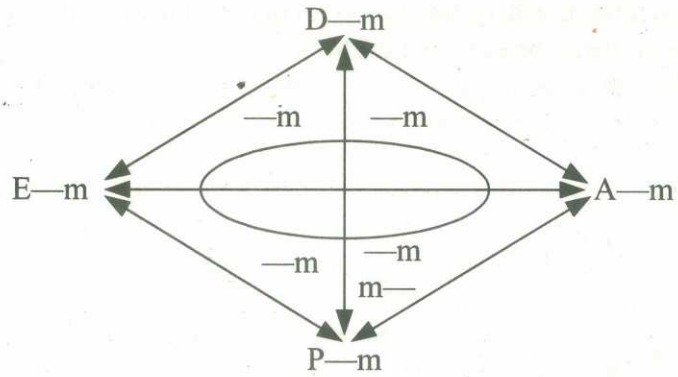


Fig. 4.4 : Deep structure of SIL

Annotation

It appears from the above figure that any specific substantive idea (and also the term denoting it) used to formulate a name-of-subject belongs to any one of the following elementary categories (D, E, A, P) and modifier:

1) *D=Discipline*: An elementary category that includes the conventional field of study, or any aggregate of such fields, or artificially created fields analogous to those mentioned above; e.g. Physics, Biotechnology, Ocean science, Library and Information Science, etc.

2) *E=Entity*: An elementary category that includes manifestations having perceptual correlates, or only conceptual existence, as contrasted with their properties, and actions performed by them or on them; e.g. Energy, Light, Plants, Animals, Place, Time, Environment, etc.

3) *A=Action*: An elementary category that includes manifestations denoting the concept of 'doing'. An action may manifest itself as Self Action or External Action. For examples: Function, Migration, etc. are Self Actions; and Treatment, Selection, organization, and Evaluation, etc. are External Actions.

4) *P=Property*: An elementary category that includes manifestations denoting the concept of 'attribute'—qualitative or quantitative; e.g. Property, Effect, Power, Capability, Efficiency, Utility, Form, etc.

5) *M=Modifier: The term 'm*odifier' refers to an idea used to qualify the manifestation of any one the elementary categories D, E, A and P, without disturbing its conceptual wholeness. As a result, it decreases the extension and increases the intension of the qualified manifestation without disturbing its conceptual wholeness. A modifier can modify a manifestation of any one of the elementary categories, as well as two or more elementary categories. Any manifestation of any elementary category may serve as the basis for deriving a Modifier. Modifiers generally create Species/Types. Modifiers are of two types:

   a) *Common Modifiers:* They refer to Space (e.g. Indian Music), Time (e.g. 15th Century Drama), Environment (e.g. Desert Birds), and Form (e.g. Dictionary of Physics). Common modifiers have the property of modifying a combination of two or more elementary categories.

   b) *Special Modifiers:* A special modifier is used to modify only one of the elementary categories. That is, it may be of Discipline-based, or Entity-based, or Property-based, or Action-based. Special modifiers can be grouped into two types:

      i) those that require a phrase or auxiliary words to be inserted between the term and thus forming a complex phrase, e.g. Indexing using computers; and

ii) those that do not require auxiliary words or phrase to be inserted in between the terms, but automatically form an acceptable compound term denoting Species/Type, e.g. 'Infectious' in 'Infectious diseases'.

Some other working concepts associated with the GT-SIL are explained below:

Organising Classification and Associative Classification

According to the GT-SIL, classification is a combination of both organising classification and associative classification. In other words, an indexing system is a combination of both organising classification and associative classification. The tasks involved in creating an organising classification are the categorisation of concepts and their organisation with the help of the DS-SILs. In organising classification compound subjects are based on genus-species, whole-part, and other inter-facet relationships. Here, classification is used to distinguish and rank each subject from all other subjects with reference to its COordinate—Superordinate—Subordinate—COllateral (COSSCO) relationships. It is achieved by means of interpolation and extrapolation of successive superordinates of each manifestation. The result of organizing classification is always a hierarchy.

In associative classification, a subject is distinguished from other subjects based on the reference of how it is associated with other subjects without reference to its COSSCO relationships. The result of associative classification is always a relative index.

Base and Core

In the context of constructing compound subject propositions, when the purpose is to bring together all or major portion of information relating to a particular manifestation or manifestations of a particular elementary category, as the case may be, the manifestation/category is Base. In a complex-subject proposition, any one of the subjects can be decided to be the Base subject depending upon the purpose in hand. For example: for a document on 'Eye cancer', 'Eye' is the Base subject in an Eye Hospital Library, and 'Cancer' is to be considered as the Base subject for a Cancer Research Centre.

When the purpose is to bring together within a recognized Base, all or major portion of information pertaining to manifestations of one or more elementary categories, the category or categories concerned may be deemed to be the Core of the concerned Base. Core lies within the Base, and which one will be the Base or Core depends on the collection or purpose of the library. For example: In DDC, 'Medicine' is the Base, and the 'Human body' and its 'Organs' constitute the Core of the Base.

Postulates Associated with the DS-SILs

The DS-SILs and its following associated postulates may be considered as the essence of the GT-SIL:

1) *Basic Sequence:* The sequence D followed by E (Either modified or unmodified) appropriately interpolated or extrapolated by A and P (either modified or unmodified) is a logical sequence of the elements of a Basic Chain manifesting in a compound subject-proposition. Any A or P may have A and/or P directly related to it. Their position is always after the A or P to which they are related. This sequence may be used as the basis for generating organising classification; and which in turn may form the basis of associative classifications. The logic behind this proposition is based on the purpose of generating a "General to Specific" sequence in the vertical arrangement of subject propositions.

2) *Source Organising Classification:* The basic sequence of manifestations augmented by the interpolation and extrapolation of successive superordinates of each manifestation, whenever warranted, gives rise to a Basic Modulated Chain; which can generate a source organising classification in alphabetical arrangement with the aid of suitable of apparatus introduce for this purpose. The basic modulated chain can be manipulated to generate chains meant for different organising classifications, and different associative classifications. This implies that there is no single absolute version of organising or associative classification. Classification is always purpose-oriented that determines the optimally efficient and effective versions of classification. The common organising structure of the manipulated versions of basic modulated chains constitutes the surface structure of specific SIL.

3) *Inter-facet relationship in Basic Chains:* In basic chains, the relationship between any two manifestations is the inter-facet relationship. For example, in the basic chain *"Agri-culture, Rice, Harvesting",* the relationship between any two manifestations is inter-facet relationship.

4) *Modifyee-Modified Relationship:* In basic modulated chains, the relationship between the modifyee and the modified manifestation is the Genus-Species relationship. For example, the relationship between "Disease" and "Infectious disease" is of this type.

5) *Associative Classification Effect:* The simple cycle permutation of each of the sought terms, in any style, with the indication of structure of subject propositions meant for organising classification has every effect of an associative classification.

6) *Systematic Grouping:* Only the notational representations of modulated chains can ensure in arrangement the ideal systematic grouping by juxtaposition. Only the alphabetical arrangement of modulated chains with suitable notational apparatus is the closest approximation to the purely notational grouping. Grouping by referencing has the organising classification effect.

7) *Synonyms, quasi-synonyms, and antonyms* can be controlled only by referencing, or by some acceptable substitute for referencing.

Characteristic Features of POPSI-Basic

The DS-SILs and its associated postulates are to be used as the basis for designing a Basic SIL, and it has been called POPSI-Basic. The postulates, principles and the working concepts of POPSI are the same as those forming part of the General Theory of SIL. POPSI as a process or operation for preparing subject-propositions consists primarily of (a) Analysis; (b) Synthesis; and (c) Permutation. The work of analysis and synthesis is primarily based on the postulates about the DS-SILs, and the work of permutation is based on the cyclic permutation of each term-of approach, either individually or in association with other terms for generating associative classification effect in alphabetical arrangement. The task of analysis and synthesis for POPSI–Basic is largely guided by the following POPSI-table. This table prescribes the notations that can be used to indicate mostly the categories of manifestation as well as the position of manifestations in subject-propositions. This feature of POPSI table is essential for mechanizing the horizontal arrangement of manifestation in subject-propositions, as well as for the vertical alphabetisation leading to organising classification.

Table 4.2: POPSI Table

0	Form Modifier
1	General Treatment
2	Phase relation 2.1 General 2.2 Bias 2.3 Comparison 2.4 Similarity 2.5 Difference 2.6 Application 2.7 Influence
	Common Modifiers
3	Time modifier
4	Environment modifier
5	Place modifier
6	Entity (E)
7	Discipline (D)
	. 1 Action (A)
	. 2 Property (P)
	, Part . Species/Type — Special modifier
	Preceded by the notation of the manifestation in relation to which number is to be it is related
	Note: A & P can go with another A&P also. In that case, the Action/Property. Preceded by that of its manifestation.
7	Core (C)
8	Base (B)
	Features analogous to 6 Entity / 7 Discipline / Action / Property.

Principles of Sequence for POPSI-Basic

The logical sequence of manifestations prescribed for subject-propositions meant for POPSI-Basic is the Basic sequence as prescribed in the postulates associated with the DS-SILs. A Species (Types)/Part follows immediately the manifestation in relation to which it is a Species (Type)/Part. A modifier (m) follows immediately the manifestation in relation to which it is a Modifier (m). The notation to be assigned to the manifestation of Discipline (D) in POPSI-Basic is always omitted .The ordinal value of this omitted notation is always taken to be the highest. The same procedure is adopted in case of any Base (B), whether it is a manifestation of D or not.

In subject-propositions for organizing classification, the different manifestations are normally arranged horizontally in the decreasing sequence of the ordinal values of the notations indicating their respective categories.

When arranged horizontally in the decreasing sequence of their ordinal values, the punctuation marks hyphen (-), full stop (.), and comma (,) fall in the same sequence. The horizontal sequence of the manifestations Modifier, Species/Type, and Part in a subject-proposition for organizing classification is the same as that mentioned above. In the vertical sequence of subject-propositions their ordinal values fall in the following increasing sequence: comma (,), full-stop (.), and hyphen(-).

Steps in POPSI-Basic

1) Analysis of the subject-proposition (=Analysis): It consists of identifying the elementary categories and modifiers in the statement of the subject.

2) Formalisation of the Subject-Proposition (=Formulation): It consists of the formalisation of the subject-proposition on the basis of the results of the step 1 (Analysis) according to the principles of sequence of components indicating the status of each facet.

3) Standardisation of the Subject-proposition (=Standardisation): It consists of deciding the standard term especially for the manifestation having synonyms.

4) Modulation of the Subject-Proposition (=Modulation): It consists of augmenting the standardized subject proposition by interpolating and extrapolating, as the case may be, the successive superordinates of each manifestation by using standard terms with indication of their synonyms, if any. [*Note: A* Classaurus is the tool to guide the operation in steps 3 and 4 with assurance of consistency in practice].

5) Preparation of the Entry for Organising Classification (=Preparation of EOC): It consists of preparing the entry for generating organising classification by juxtaposition of entries in alphabetical sequence. [*Note*: The operation in this step is guided by the above mentioned POPSI-table].

6) Decision about Terms-of –Approach (=Decision about TA): It consists of deciding the terms-of-approach for generating associative classification for its effects; and of controlling synonyms.

7) Preparation of Entries of Associative Classification (=Preparation of EAC): It consists of preparation of entries under each term-of-approach, their individually, or in association with other terms, by cyclic permutation of sought terms for generating associative classification effect in alphabetical arrangement in such a way that under each term, a more or less systematic arrangement of subject-proposition is found.

8) Alphabetical Arrangement of Entries (=Alphabetisation): It consists of arranging all the entries in alphabetical sequence.

Demonstration of the Procedure for the POPSI-Basic

0 *Subject Indicative Expression*

Implications of technologies changes in rice cultivation for India

1) *Analysis*

D=Agriculture

E=Rice

A to E =Cultivation

P of A to E =Technology

P of P of A to E=Changes

P of P of P of A to E =Implications

m of the whole subject proposition=India

2) *Formalisation*

Agriculture (D), Rice (E), Cultivation (A to E), Technology (P of A to E), Changes (P of P of A to E), Implications (P of P of P of A to E), India (m of the whole subject proposition).

3) *Standardisation*

Agriculture (D), Field crop. Rice (E), Cultivation (A to E), Technology (P of A to E), Change (P of P of A to E), Implication (P of P of P of A to E), India (m of the whole subject proposition).

4) *Modulation*

Agriculture (D), Field Crop. Rice (E), Cultivation (A to E), Technology (P of A to E), Change (P of P of A to E), Implication (P of P of P of A to E), Asia, India (m of the whole subject proposition).

5) *Preparation of EOC*

Agriculture 6 Field Crop. Rice 6.1 Cultivation 6.1.2 Technology 6.1.2.2 Change 6.1.2.2.2 Implication 5 (for) Asia, India

6) *Decision about EAC*

Use each term other than 'Agriculture', 'Change', and 'Implication' as term-of –approach.

7) *Preparation of EAC*

Field Crop

Agriculture 6 Field Crop. Rice 6.1Cultivation 6.1.2 Technology 6.1.2.2 Change 6.1.2.2.2 Implication 5 (for) Asia, India

[*Note:* Similarly, EAC is to be prepared under each of the other terms-of–approach. Alternatively, making it multifaceted by using terms from upper links especially when warranted can make the heading of an entry for associative classification more specific. For example, the heading with 'Technology' as the term-of-approach may, in this case, consists of the following: "Technology, Cultivation, Rice"].

8) *Alphabetisation*

Arrange the entries according to the alphabetical order by terms-of-approach.

POPSI-Specific

As the situation prevails today it has become imperative for each and every information centre including a library to design and develop its own SIL in order to satisfy the condition of achieving its own specific objective. There has always been a tradition to depend upon a designer of a SIL. It is always found to be inadequate to satisfy the need at the local level. No standard SIL has yet been found to be suitable and adequate to meet the specific requirements of depth indexing in an area of micro subject. The basic assumption leading to the development of POPSI-Specific is that subject indexing is always a specific requirement-oriented activity. Difference in requirement would call for difference in syntax of the subject proposition; flexibility should be its rule, not the rigidity. Based on this assumption, POPSI tries to find out what is logically basic, and amenable to systematic manipulation to meet specific requirement. The POPSI-Basic is a product of the application of the GT-SIL and it is readily amenable to the systematic manipulation to generate purpose-oriented oriented specific versions known as POPSI-Specific. POPSI-Specific is always a derivation from the POPSI-Basic according to special decisions and rules to meet specific requirements at the local level. It may be noted here that this approach is totally different from that of earlier contributors of different SILs.

Principles of Sequence for POPSI-Specific

The postulates, principles and working concepts of POPSI–Specific are the same as those of POPSI–Basic forming the part of the GT-SIL. When the Base (B) and Core (C) do not correspond to those of POPSI–Basic, the guidance about the sequence of manifestations is readily available from the POPSI–Basic itself. The general principle in this case would be as follows: B (either modified or unmodified), is followed by C (either modified or unmodified), and the rest of the manifestations in the sequence as suggested in the postulates of basic sequence associated with the DS-SILs forming part of the GT-SIL.

Demonstration of the Procedure for POPSI-Specific

Let us consider the earlier example. Suppose that the purpose is to bring all or major portion of information pertaining to 'Cultivation Technology' together irrespective of its application to specific crops. In such a situation, 'Cultivation', which is a manifestation of Action (A), and "Technology", which is a manifestation of Property to Action (P to A), according to POPSI-Basic, are to be fused together, and the resulting manifestation 'Cultivation Technology' is to be treated as the Base (B). Suppose further that within the Base 'Cultivation Technology' we intend to organize information by 'Soil Preparation Technology', 'Sowing Technology', 'Weeding Technology', 'Harvesting Technology', etc. In that case they are to be treated as Sub-Base of the Base 'Cultivation Technology'.

Suppose still further that we intend to organise all these technologies by specific crops. In that case, the sub-bases are to be sharpened further by using each type of crop and specific crop as a Modifier of the Base and the Sub-Bases. When these are the primary decisions, the resulting indexing is called 'POPSI-Specific'; and it is implemented by systematic manipulation of POPSI-Basic. It will be evident from the following demonstration:

0 *Subject Indicative Expression*

Implication of technological changes in rice cultivation for India.

1 *Analysis*

Base (B) = Cultivation Technology

Modifier of the Base (m of B) = Rice

Property (P) of the (extended) B (P to B) = Changes

P of P of B = Implication

m of the whole subject proposition = India

2 *Formalisation*

Cultivation Technology—Rice (Extended B), Changes (P of B), Implication (P of P of B), India (m of the whole subject proportion).

3 *Standardisation*

Cultivation Technology—Rice (B), Change (P of B), Implication (P of P of B), India (m of the whole subject proposition).

4 *Modulation*

Cultivation technology—Field crop. Rice (B), Change (P of B), Implication (P of P of B), Asia. India (m of the whole subject proposition)

5 *Preparation of EOC*

Cultivation technology—Field crop. Rice. 9.2 Change 9.2.2 Implication 5 (for) Asia, India

6 *Decision of TA*

Use each term other than 'Cultivation technology', 'Change', and 'Implication' as a term of approach.

7 *Prèparation of EAC*

Field Crop

Cultivation technology—Field crop. Rice 9.2 Change 9.2.2 Implication 5 (for) Asia, India

[Similarly under each of the other sought terms-of-approach, EAC is to be prepared].

Alternative multifaceted heading:

Field crop, Cultivation Technology

8 *Alphabetisation*

Arrange the entries according to the alphabetical order by terms-of-approach.

[Note: Evidently; this POPSI-Specific has been derived from POPSI-Basic by a systematic manipulation based on the following decisions:

a) Elimination of the discipline term "Agriculture"; and

b) Recognition of "Cultivation Technology" as the Base (B)].

Self Check Exercises

3) "Kaiser started from the point where Cutter left"—Discuss.

4) Mention the different types of relationships and their respective relational operators as proposed J.E.L. Farradane in his Relational Indexing.

5) How are the syntactical and semantic relationships dealt with in PRECIS?

6) Discuss the basic assumptions that led to the development of POPSI-Specific.

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

.....

.....

.....

.....

.....

.....

.....

4.4 POST-COORDINATE INDEXING SYSTEMS

Basically all the indexing methods are, by nature, coordinate indexing. The purpose of using a combination or coordination of component terms is to describe the contents of the documents more accurately. In case of pre-coordinate indexing system, coordination of component terms is carried out at the time of indexing (i.e. at input stage) in anticipation of the users' approach. In such a system, the most important aspect is to determine the order of significance or citation order by following the syntactical rules of the given indexing language. Rigidity of the significance order may not meet all the approaches of users. To satisfy the varieties of approaches of all the users, there is the provision for multiple entries

by rotating or cycling of the component terms representing the subject of the document. It has been observed that even the acceptance of multiple entry system covers only a fraction of the possible number of the total permutation, which in turn, results into the failure of the index file to provide a particular pattern of combination which the user is looking for. Consequently, a large portion of probable approach points is left uncovered. In pre-coordinate indexing system, both the indexer and the searcher are required to understand the mechanism of the system—the indexer for arriving at the most preferred citation order and the searcher for formulating an appropriate search strategy—in order to achieve the highest possible degree of matching of concepts. Pre-coordinate indexing system is non-manipulative. The searcher has no choice but to follow the citation order specified by the indexer. The above considerations and difficulties stemming from the pre-coordination of terms led to the development of post-coordinate indexing system, sometimes referred to as Coordinate Indexing System.

In post-coordinate indexing system, component concepts (denoted by the terms) of a subject are kept separately uncoordinated by the indexer, and the user does the coordination of concepts at the time of searching (i.e. at the output stage). The searcher has wide options for the free manipulation of the classes at the time of searching in order to achieve whatever logical operations are required. As the coordination of concepts are done at the output (search) stage this is known as Post Coordinate Indexing System. The Post-coordinate indexing system is of two types:

Term Entry System: Here, index entries for a document are made under each of the appropriate subject headings and these entries are filed according to alphabetical order. Under this system, the number of index entries for a document is dependent on the number of component terms associated with the thought content of the document. Here, terms are posted on the item. Searching of two files (Term Profile and Document Profile) is required in this system. For examples, Uniterm, Peek-a-boo, etc.

Item Entry System: It takes the opposite approach to term entry system and prepares a single entry for each document (item), using a physical form, which permits access to the entry from all appropriate headings. Here, items are posted on the term. Item entry system involves the searching of one file (i.e. Term Profile) only. For example, Edge-notched card.

4.4.1 Uniterm

Martimer Taube devised the Uniterm indexing system in 1953 to organise a collection of documents at the Armed Services Technical Information Agency (ASTIA) of Atomic Energy Commission, Washington. The system is based on concept coordination, where each component term (uniterm) is independent of all other terms and serves as a unique autonomous access point to all relevant items in the collection. Uniterm indexing system had a number of distinctive characteristics:

- Indexing by single words only;
- Terms are extracted from the text of the document indexed;
- No control over those terms;
- Indexing, being reduced to word extraction, can be conducted by relatively low-level personnel.

However, immediately after its development it was realized that concepts can be represented always by a single term. As a consequence it can be herself switched over to Unit concepts rather than Unit form. So sometimes this is also referred to as Unit Concept Indexing.

Here each document is assigned an accession number or 'address'. The thought content or subject of the document is analysed into uniterms. For each document an abstract is prepared and is recorded on the card along with the selected uniterms. The accession number assigned to a document is also written on the card. The card also contains bibliographical details of the document. These cards are arranged according to the ascending order of the accession numbers of the documents and are called *Numerical File.*

For each uniterm, a term card is prepared and it is divided into 10 equal vertical columns (from 0 to 9) in which accession number of the document relevant to the uniterm are recorded according to the system of terminal digit posting. Terminal digit of the accession number determines the column of its posting. Term cards are arranged according to the alphabetical order. This is called the *Uniterm File* of the system.

When a search is made, the subject of the search is broken into uniterms and the pertinent uniterm cards are pulled out from the alphabetical deck of the *Uniterm File.* Uniterm cards thus pulled out are matched to find the common accession number(s). The number(s) common in all such uniterm cards represent the sum total of the component concept of the specific subject. With the help of the common accession number(s), relevant card(s) are pulled out from the *Numerical File* where full bibliographical information of the required document(s) is available.

The merit of the uniterm indexing system is its simplicity and the ease with which persons without much knowledge of subject indexing can handle it. The limitations of the system are two-fold: (a) it takes much time in searching a document because of the searching of two files—uniterm file and numerical file particularly when it is practiced manually. Due to development of technology, it is done faster now; (b) it may retrieve irrelevant documents due to false coordination of uniterms. For example, searching with the uniterms 'science' and 'philosophy' may retrieve the documents on both the subjects 'Philosophy of Science' and 'Science of Philosophy', one of which might be irrelevant to a particular user. To overcome the problem, Engineering Joint Council devised `Roles' which may be attached to the terms to resolve the ambiguity.

4.4.2 Optical Coincidence Card / Peek-a-boo

Peek-a-boo is the trade name of the optical coincidence card. It is also called 'Batten Cards'. One card may punch 500 to 10000 accession numbers depending on the size of the card and size of the ruled squares. Instead of putting accession number (as is done in uniterm system), an item is indexed by punching a hole into the appropriate position that serves to represent the document number. The following tools or equipments are required for indexing under this system:

- For punching: Hand drill (i.e. hand punch) or electrically operated model;
- For lighting: Light box—A box containing different sources of light.

An optical coincidence card is shown below:

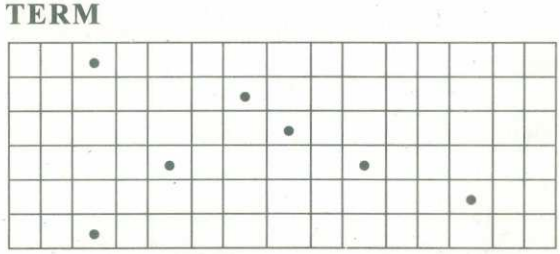


Fig. 4.5 : A sample of Optical Coincidence Card

The scanning is simplified with a beam of light that passes through the holes in the relevant cards (which are hold before the light) indicating accession number. Here one has to detect that through which hole(s) the light is passing in order to ascertain the accession number.

Advantages

1) Uniterm relies on our ability to notice the matching accession number(s) on the card(s) we are scanning having the probability of mistake. This can be avoided here.
2) They have the potential to store a relatively large number of documents.
3) It permits speedy search.

Disadvantages

1) There are chances of punching mistakes.
2) Withdrawal of card is very difficult because of the holes. Although colour cards are inserted for this purpose.
3) Like uniterm indexing system, it involves two level searches—first term profile and then documents profile.

4.4.3 Edge-Notched Card

Indexing on Edge-Notched card is based on punched card system. Their value is limited to very small collection.

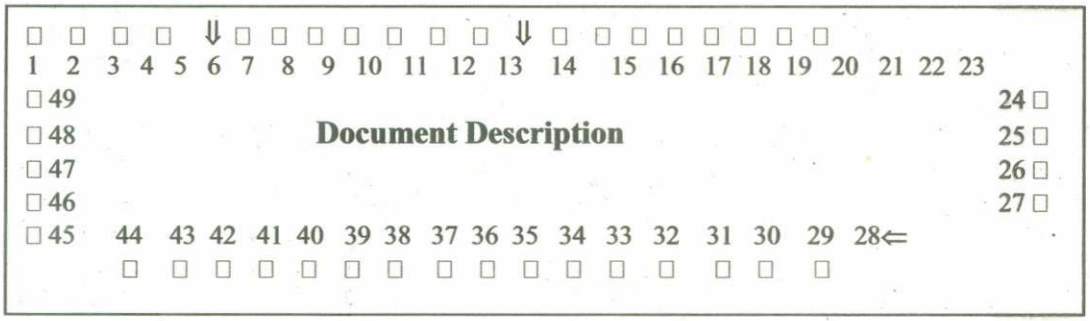
Features

1) One card corresponds to one document.
2) Number of holes on a card varies from 75 to 128.
3) A list of terms (term profile) associated with the thought content of the incoming documents is prepared along with their serial numbers.
4) Each hole represents a particular index term.
5) Subject index of a document is made by a series of notches punched out from the holes along the margin of the card.
6) The middle point of the card is used for bibliographic description of the document.
7) Equipment (physical apparatus) requires:
   - Wedge-shaped punch for notching marginal holes;
   - Needle(s) as sorting device (i.e. punched and non-punched cards)
8) Searching is carried out by inserting needle through the hole(s) corresponding to the request representation carries out searching and thereby allowing the cards punched in that position to fall.

Example: Let us consider a document on *'Chemotherapy of heart cancer among children'*. The term profile representing the serial numbers of the component terms associated with the thought content of this document appears to be as follows:

Serial No.	Terms
7	Cancer
15	Chemotherapy
28	Children

For this, one will have to punch the holes corresponding to the positions 7, 15 and 28, as shown below:



In zatacoding system of Calvin Mooers, relevant cards are sorted out by vibration after the selector needle is inserted into the pack.

Advantages

We have seen in the preceeding sections that two-level search is required in case of term entry system—i.e. searching term profile first and it is to be followed by searching the document profile. Edge-Notched Card System avoids second search.

Disadvantages

One card may contain up to a certain number of holes and as a result this system is not suitable for large collection.

These methods (4.4.2 & 4.4.3) are semi-mechanical methods and are no longer practiced.

4.4.4 Post-Coordinate Searching Devices

There are a number of devices, which are an essential part of post-coordinate search strategy to avoid the false drops, i.e. retrieval of unwanted documents because of the false coordination of terms at the time of searching. Devices used for the elimination false drops are used both for indexing and searching. They are discussed below:

a) **Use of Bound Terms**

Two or more terms in a subject may be bound to get rid of false coordination.

b) **Roles**

The role indicator is a symbol attached to the index term at the indexing stage to indicate the sense and use of the term in a particular context. Compilers of retrieval languages may develop their own roles. It may be shown that a term is functioning as 'Raw material', 'Product / Output', or that it is functioning as an 'Agent / Tool'. The role indicator is most useful in avoiding recall of terms with incorrect function. For instance, roles produced by the Engineering Joint Council (EJC) may be added to *'Alcohol'* to distinguish their functions as raw material, agent or tool, and product or output. The examples given below illustrate the use of EJC role indicators with the term *'alcohol'* in a practical situation:

Role	*Document*
1 [Raw Material / Input]	Use of *alcohol* in the production of wine
2 [Product / Output]	Manufacturing of *alcohol*
3 [Agent / Tool]	Use of *alcohol* as reagent in laboratories

c) Links

Links are special symbols used to group all the related concepts in a document separately, so that inappropriate combinations of terms are not retrieved. For example, a document number 327 on *'Welding of copper pipes and heat treatment of steel structures'* would be indexed:

WELDING	327A
COPPER	327A
PIPES	327A
HEAT TREATMENT	327B
STEEL	327B
STRUCTURES	327B

where the letters A and B indicate which terms are associated, thus ensuring that the document is not retrieved in a search for 'Welding of steel'.

d) Weighting

In weighting system, values are allocated to indexing terms or search terms according to their importance in particular documents or search programs. Weighting acts as a precision device and also as an output-ranking device. It is the method of the application of quantitative figures to the indexing / searching terms depending upon their emphasis in the document. The following table shows the weight worked out by Maron and others:

Table 4.3: Weightage System (by Maron)

Weight	Description	When Used
8/8	Major Subject	The term is highly specific and covers an entire major subject of the document.
7/8	Major Subject	The term is specific and covers most of a major subject of the document.
6/8	More Generic Subject	The term is too broad and covers a major subject.
5/8	Other Important Terms	Terms that would be used in a binary indexing system but not a major subject.
4/8	Less Generic Subject	The term relates to, but is too narrow to cover a major subject.
3/8	Minor Subject	Includes such terms as relate to results of experiments, intermediate methods, possible uses, etc.
2/8	Other Subjects	Other relevant tags.
1/8	Barely Relevant	Subject classifier would not want to use, but feels that some users might consider relevant.

The examples given below illustrate the use of weights from the above table:

Weight	*Term*	*Document*
8/8	X-Ray	Introduction to *X-Ray*
3/8	X-Ray	*X-Ray* treatments of the heart diseases

The second Cranfield Project used a less complex weighting system:

9/10 For concepts in the main theme of the document

7/8 For concepts in a major subsidiary theme

5/6 For concepts in a minor subsidiary theme

Weighting of search terms may also be a guide when adjusting search strategies. For instance, when reducing the coordination level, terms with lower weights may be excluded first, in the expectation that this will give optimum recall improvement.

Self Check Exercises

7) Distinguish between pre-coordinate and post-coordinate indexing systems?

8) What are the devices used to eliminate false drops in post-coordinate indexing?

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

.....

.....

.....

.....

.....

.....

4.5 AUTOMATIC INDEXING

In many literatures of Library and Information Science, the term 'automatic indexing' is interchangeably used with the term 'computerised indexing'. A fully automatic indexing system would be one in which indexing is conducted by computers, an internally generated thesaurus is prepared, and search strategies are developed automatically from a natural language statement of information need. Salton provides the following definition of *automatic indexing*: *When the assignment of the content identifier is carried out with the aid of modern computing equipment the operation becomes automatic indexing.* It has been suggested that the subject of a document can be derived by a mechanical analysis of the words in a document and by their arrangement in a text. In fact, all attempts at automatic indexing depend in some way or other on the test of the original document or its surrogates. The words occurring in each document are examined and substantive words are selected through statistical measurements (like word frequency calculation, total collection frequency, or frequency distribution across the documents of the collection) by the computer.

However, the use of computers in generating indexes of documents started from KWIC indexing developed by H.P. Luhn.

The idea of analysing the subject of a document through automatic counting of term occurrences was first put forward by H P Luhn of IBM in 1957. He proposed that :

a) The frequency of word occurrence in a text of the document furnishes a useful measure of word significance;

b) The relative position of a word within sentence furnishes a useful measurement for determining the significance of sentences; and

c) The significance factor of a sentence will be based on a combination of these two requirements.

The basic idea behind Luhn's automatic indexing was based on word extraction, that is, keywords were extracted from the text by counting the frequency of occurrence of words in a given document. Here, the computer was used to scan the text with the object of counting the words or phrases that occur most frequently in a machine-readable document, and the extraction programs select the words or phrases that occur most frequently to represent the subject-matter of the document. A 'stop word' list was first used to eliminate the common and non-substantive words. The system pioneered by Luhn was relatively effective and the words or phrases selected by computer were quite similar to those, which would be extracted by a human indexer.

In the early 1960s, some other attempts were made at implementing automatic indexing systems. These consisted in using the computer to scan document texts, or text excerpts such as abstracts, and in assigning as content descriptor words that occurred sufficiently frequently in a given text. A less common approach uses relative frequency in place of absolute frequency. In relative frequency approach, a word is extracted if it occurs more frequently than expected in a particular corpus. Thus in a document on 'Aerodynamics' the word 'Air Craft' and the word 'Wing' might be rejected, even though they are the most frequently occurring words in the document, and the word 'Flutter' might be selected even though, in absolute terms, it is not a high frequency words. Other approaches to automatic indexing use other types of extraction criteria in place of, or along with the statistical criteria, word position in the document, word type, or even the emphasis placed on words in printing—(e.g. boldface and italics)—may all be used as the basis for selection. Subsequently linguistics led the way by pointing out that a number of linguistic processes were essential for the generation of effective content identifiers characterizing natural language texts.

An ideal computerised indexing is one that has the ability to create and modify new subject terms mechanically, by minimising or without the help of human intellectual efforts. As computer can understand only machine code, so it is necessary to translate the information into machine code and in a fixed machine-readable format. Usually, the titles and abstracts are used for the purpose of computerised indexing. However there are two assumptions:

a) There is a collection of documents; each contains information on one or several subjects.

b) There exists a set of index terms or categories from which one or several of them can describe/represent the subject content of every document in the collection.

4.5.1 Manual Indexing vs. Computerised Indexing

Manual Indexing	Computerised Indexing
1) Identifying and selecting keywords from the tile, abstract and full text of the document to represent its content.	1) Keywords and/or phrases denoting the subject matter of the document are extracted only from the title and abstract rather than the document's full text.
2) Content analysis of the document is purely a mental process and carried out by the human indexer.	2) The computer does content analysis by following the human instructions in the form of a computer programming.
3) Human indexer makes inferences and judgment in selecting index terms judiciously.	3) Computer cannot think and draw inferences like human indexer and as such, it can select or match keywords, which are provided as input text.
4) Human indexer selects and excludes index terms on the basis of semantic, syntactical as well as contextual considerations.	4) It is possible to instruct a computer through proper programming to select, or exclude a term by following the rules of semantic, syntactical and contextual connotations, like human indexer.

(Table contd.)

Manual Indexing	Computerised Indexing
5) Scanning, analyzing the critical views, understanding the concepts and using indexer's own subject knowledge and previous experience do indexing.	5) Computer cannot do this. It involves less intellectual effort.
6) Selected index terms less in number.	6) Selected index terms are more in number.
7) It is time consuming.	7) It takes less time.
8) It is expensive.	8) Index entries can be produced at lower cost.
9) It is very difficult to maintain consistency in indexing.	9) Consistency in indexing is maintained.

4.5.2 Methods of Computerised Indexing

4.5.2.1 Keyword Indexing

An indexing system without controlling the vocabulary may be referred as 'Natural Language Indexing' or sometimes as 'Free Text Indexing'. Keyword indexing is also known as Natural Language or Free Text Indexing. 'Keyword' means catch word or significant word or subject denoting word taken mainly from the titles and / or sometimes from abstract or text of the document for the purpose of indexing. Thus keyword indexing is based on the natural language of the documents to generate index entries and no controlled vocabulary is required for this indexing system. Keyword indexing is not new. It existed in the nineteenth century, when it was referred to as a 'catchword indexing'. Computers began to be used to aid information retrieval system in the 1950s. The Central Intelligence Agency (CIA) of USA is said to be the first organization to use the machine-produced keywords index from Title since 1952. H P Luhn and his associates produced and distributed copies of machine produced permuted title indexes in the International Conference of Scientific Information held at Washington in 1958, which he named it as Keyword-In-Context (KWIC) index and reported the method of generation of KWIC index in a paper. American Chemical Society established the value of KWIC after its adoption in 1961 for its publication 'Chemical Titles':

KWIC (Keyword-In-Context) Index

As told earlier, H P Luhn is credited for the development of KWIC index. This index was based on the keywords in the title of a paper and was produced with the help of computers. Each entry in KWIC index consists of following three parts:

a) *Keywords*: Significant or subject denoting words which serve as approach terms;

b) *Context:* Keywords selected also specify the particular context of the document (i.e. usually the rest of the terms of the title).

c) *Identification or Location Code*: Code used (usually the serial numbers of the entries in the main part) to provide address of the document where full bibliographic description of the document will be available.

The operational stages of KWIC indexing consist of the following:

a) Mark the significant words or prepare the 'stop list' and keep it in computer. The 'stop list' refers to a list of words, which are considered to have no value for indexing / retrieval. These may include insignificant words like articles (a, an, the), prepositions, conjunctions, pronouns, auxiliary verbs together with such general words as 'aspect', 'different', 'very', etc. Each major search system has defined its own 'stop list' ;

b) Selection of keywords from the title and / or abstract and / or full text of the document excluding the stop words;

c) KWIC routine serves to rotate the title to make it accessible from each significant term. In view of this, manipulate the title or title like phrase in such a way that each keyword serves as the approach term and comes in the beginning (or in the middle) by rotation followed by rest of the title;

d) Separate the last word and first word of the title by using a symbol say, stroke [/] (sometime an asterisk '*' is used) in an entry. Keywords are usually printed in bold type face;

e) Put the identification / location code at the right end of each entry; and finally

f) Arrange the entries alphabetically by keywords.

Let us take the title 'control of damages of rice by insets' to demonstrate the index entries generated through KWIC principle:

Control of damages of rice by insets 118

Damages of rice by insets / Control of 118

Insets / Control of damages of rice by 118

Rice by insets / Control of damages of 118

In the computer generated index, the keywords can be positioned at centre also.

Variations of KWIC

Two important other versions of keyword index are KWOC and KWAC, which are discussed below:

KWOC (key-word out-of-context) Index

The KWOC is a variant of KWIC index. Here, each keyword is taken out and printed separately in the left hand margin with the complete title in its normal order printed to the right. For examples,

Control	Control of damages of rice by insets	118
Damages	Control of damages of rice by insets	118
Insets	Control of damages of rice by insets	118
Rice	Control of damages of rice by insets	118

Sometime, keyword is printed as heading and the title is printed in the next line instead of the same line as shown above. For examples,

Control
Control of damages of rice by insets 118

Damages
Control of damages of rice by insets 118

Insets
Control of damages of rice by insets 118

Rice
Control of damages of rice by insets 118

KWAC (key-word Augmented-in-context) Index

KWAC also stands for 'key-word-and-context'. In many cases, title cannot always represent the thought content of the document co-extensively. KWIC and KWOC could not solve the problem of the retrieval of irrelevant document. In order to solve the problem of false drops, KWAC provides the enrichment of the keywords of the title with additional keywords taken either from the abstract or from the original text of the document and are inserted into the title or added at the end

to give further index entries. KWAC is also called enriched KWIC or KWOC. CBAC (Chemical Biological Activities) of BIOSIS uses KWAC index where title is enriched by another title like phrase formulated by the indexer.

Other Versions

A number of varieties of keyword index are noticed in the literature and they differ only in terms of their formats but indexing techniques and principle remain more or less same. They are

i) *KWWC (Key-Word-With-Context) Index*, where only the part of the title (instead of full title) relevant to the keyword is considered as entry term.

ii) *KEYTALPHA (Key-Term Alphabetical) Index*. It is permuted subject index that lists only keywords assigned to each abstract. Keytalpha index is being used in the 'Oceanic Abstract'.

iii) *WADEX (Word and Author Index)*. It is an improved version of KWIC index where the names of authors are also treated as keyword in addition to the significant subject term and thus facilitates to satisfy author approach of the documents also. It is used in 'Applied Mechanics Review'. AKWIC (Author and keyword in context) index is another version of WADEX.

iv) *DKWTC (Double KWIC) Index.* It is another improved version of KWIC index.

v) *KLIC (Key-Letter-In-Context) Index.* This system allows truncation of word (instead of complete word), either at the beginning (i.e. left truncation) or at the end (i.e. right truncation), where a fragment (i.e. key letters) can be specified and the computer will pick up any term containing that fragment. The Chemical Society (London) published a KLIC index as a guide to truncation. The KLIC index indicates which terms any particular word fragment will capture.

Uses of Keyword Index

A number of indexing and abstracting services prepare their subject indexes by using keyword indexing techniques. They are nothing but the variations of keyword indexing apart from those mentioned above. Some notable examples are:

- Chemical Titles;
- BASIC (Biological Abstracts Subject In Context);
- Keyword Index of Chemical Abstracts;
- CBAC (Chemical Biological Activities);
- KWIT (Keyword-In-Title) of Laurence Burkeley Laboratory;
- SWIFT (Selected Words in Full Titles); and
- SAPIR (System of Automatic Processing and Indexing of Reports).

Advantages

1) The principal merit of keyword indexing is the speed with which it can be produced;

2) The production of keyword index does not involve trained indexing staff. What is required is an expressive title coextensive to the specific subject of the document;

3) Involves minimum intellectual effort;

4) Vocabulary control need not be used; and

5) Satisfies the current approaches of users.

Disadvantages

1) Most of the terms used in science and technology are standardized, but the situation is different in case of Humanities and Social Sciences. Since no controlled vocabulary is used, keyword indexing appears to be unsatisfactory for the subjects of Humanities and Social Sciences;

2) Related topics are scattered. The efficiency of keyword indexing is invariably the question of reliability of expressive title of document as most such indexes are based on titles. If the title is not representative the system will become ineffective, particularly in Humanities and Social Science subjects;

3) Search of a topic may have to be done under several keywords;

4) Search time is high;

5) Searchers very often lead to high recall and low precision; and

6) Fails to meet the exhaustive approach for a large collection.

Self Check Exercise

9) What are the different versions of keyword indexing?

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

.....

.....

.....

.....

.....

4.5.2.2 Other Methods of Automatic Indexing

Since the KWIC indexing methods various methods generating automatic indexes have been tried. In fact, all attempts at computerized indexing were based on two basic methods: Statistical analysis; and Syntactic and semantic analysis. These are discussed bellow:

a) **Statistical Analysis**

The statistical analysis methods are based on the hypothesis that occurrence of a word in the text indicates its importance. On the basis of this hypothesis a prediction can be made about the subject terms that can be assigned to the document. The computer program can list all the words in a document. The words are grouped by number of occurrences and arranged alphabetically within each frequency. Generally articles, conjunctions, prepositions and pronouns are excluded using a 'stop list' file. Words having same stem can be counted either as the same or as different words. The following methods are adopted in measuring the word significance:

a) *Weighting by location*

A word appearing in the title might be assigned a greater weight than a word appearing in the body of the work.

b) *Relative frequency weighting*

This is based upon the relation between the number of times the words is used in the document being indexed and the number of times the same word appears in the sample of other documents.

c) *Use of noun phrase*

Noun and adjective noun phrases can be selected as index terms and these are selected from the title or abstract of the document.

d) *Use of thesaurus*

A thesaurus can be used to control synonyms and otherwise related terms. In this way, the count of some word types increases as is the separation between 'good' and 'poor' index terms.

e) *Use of association factor*

By means of statistical association and correlation techniques, the degree of term relatedness, that is, the likelihood that that two terms will appear in the same document, is computed and used for selecting index terms.

f) *Maximum-depth indexing*

This procedure indexes a document by all of its content words and weights these words, if desired, by the number of occurrences in the document. In this way, the problem of selecting term is avoided.

b) **Syntactic and Semantic Analysis**

Among the linguistic techniques of interest, the syntactical and semantic analyses are most important in the development of information analysis system needed for computerised indexing. According to Salton, most information analysis systems are based on the recognition of certain key elements, often chosen from a pre-constructed list of acceptable terms, and on the determination of rules by which these basic elements are combined into larger units. The syntactical analysis identifies the role of the word in the sentence, that is, its grammatical class (i.e. parts of speech) and relation among words in the sentence. Whereas semantic analysis helps to establish the paradigmatic or class relations among terms so as to associate words with simple concepts. The main objective of semantic analysis is to identify subject and content bearing words of the document or surrogate text.

Among the linguistic techniques of interest, the following were considered to be of significant:

a) Use of hierarchical term arrangements relating to the content terms in the given subject area can help to expand the standard content description by adding superordinate and/or subordinate terms to a given content description.

b) Use of synonym dictionaries or thesauri can help to broaden the original context description through a complete class of related terms.

c) Use of syntactical analysis systems capable of specifying syntactic roles of each term and of forming complex content descriptions consisting of term phrases and large syntactic units. A syntactic analysis scheme makes it possible to supply specific content identification.

Use of semantic analysis systems in which the syntactic units are supplemented by semantic roles attach to the entities making up a given content description. Semantic analysis systems utilise various kinds of knowledge extraneous to the documents, often specified by pre-constructed 'semantic graphs' and other related constructs.

Advanced linguistic techniques, that is, the application of computer to analyse the structure and meaning of language led by Noam Chomsky. The linguistic model proposed by Chomsky distinguishes between surface structure and deep structure of a language. By means of transformational grammar, a structure can go through a series of transformations that will exhibit the deep structure. Chomsky found that a purely syntactic transformation could provide a semantic interpretation of the sentence.

Advantages

The advantages of computerized indexing are manifold like level of consistency in indexing can be maintained; index entries can be produced at a lower cost in the long run; indexing time can be reduced; and better retrieval effectiveness can be achieved.

Disadvantages

The main criticism against computerized indexing centres round the fact that a term occurs several times in a document may not always be regarded as a significant term.

4.5.3 File Organisation

Organisation of files in manually operated libraries are expensive, time consuming, labourious and error-prone. Moreover, manual organisation of data/files often leads to duplication of data or data redundancy. Various files maintained in traditional library system are sequential in nature. For example, catalogue *cards* are arranged in order of search keys (author, title, subject, series etc.) in the form of access points transcribed at the top of the cards.

In computerised indexing system, data elements are stored in suitable digital media and it is possible to manipulate, retrieve and view data elements quite easily. The efficiency of such computerised indexing system largely depends on its file structure.

File organisation forms an important element in computerised indexing. File organization is a technique for physically arranging the records of a file on secondary storage devices. This is the technique for organisation of the data of a file into records, blocks and access structures.

A file contains data that is required for information processing. These data are about entities. An *entity* is anything about which data can be stored (e.g. book). The essential properties of an entity are called *attributes* (e.g. author, title, edition, etc. are attributes of the entity book). Each attribute of an entity is represented in storage by a data item. A data item is assigned a name in order to refer to it in storage, processing and retrieval. Data items are usually grouped together to describe an entity. The data representation in storage of each instance of an entity is commonly called as a *record.* A collection of related records is called a *file.*

When data are stored on auxiliary storage devices (e.g. hard disk), the chosen method of file organization will determine how the data can be accessed. The organization of data in a file is influenced by a number of factors, but the most important among them is the time required to access a record and to transfer the data to the primary storage or to write a record or to modify a record.

A schematic view of file organisation techniques is represented in figure 4.6.

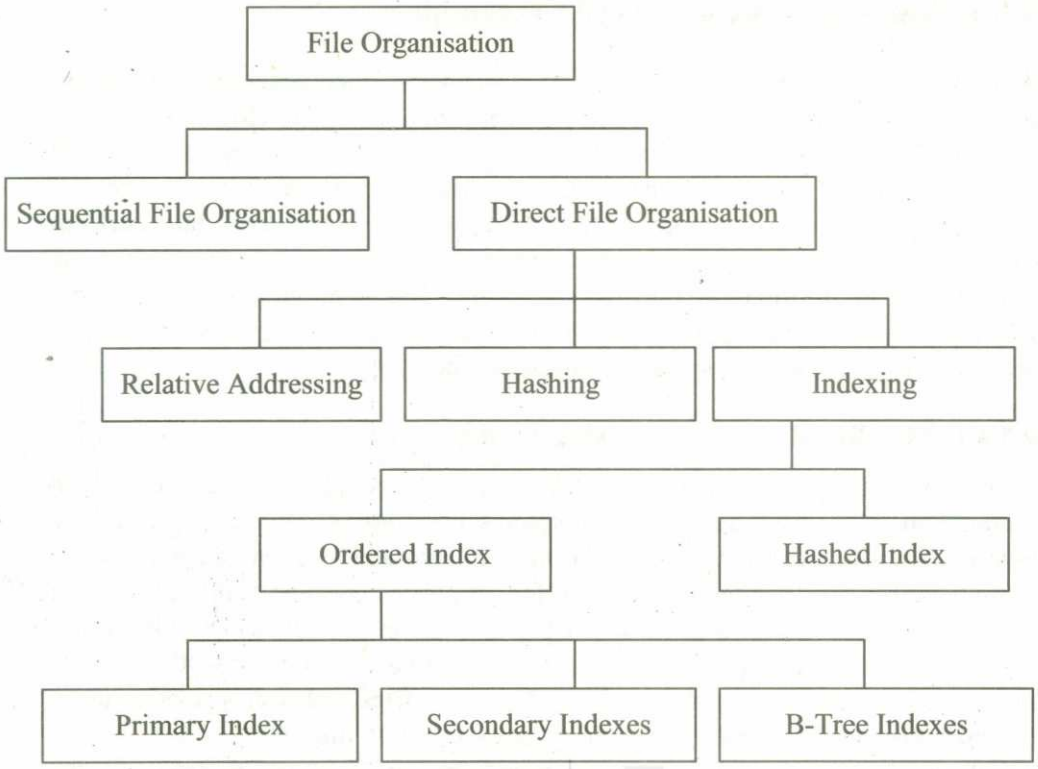


Fig. 4.6: Schematic view of file organisation

1) *Sequential File Organisation:* In this technique, records are stored in some predetermined sequence, one after another. One field, referred to as the primary key, usually determines their sequence of order.

2) *Direct File Organisation:* This technique supports direct access (also called random access), in which records can be accessed instantaneously and in any order from the data scattered throughout the disk.

   2.1 *Relative Addressing:* It is the simplest method of finding a record. Here, a record's primary key is associated with a specific physical storage location and contents of the records are stored in this address.

   2.2 *Hashing:* It is a method for determining the physical location of a record. Here, the record key is processed mathematically, and another number is computed that represents the location where the record will be stored.

   2.3 *Indexing:* It is a procedure for locating a record in a file stored randomly throughout the disk. Here, a primary index associates a primary key with the physical location in which a record is stored.

      2.3.1 *Ordered Index:* It is based on a sorted ordering of values.

         2.3.1.1 *Primary Index:* The records in the indexed file can be stored in some sorted order. If the file containing the records is sequentially ordered, the index whose search key specifies the sequential order of the file is called the primary index.

         2.3.1.2 *Secondary Index:* Indexes whose search key specifies an order that is different from the sequential order of the file are called secondary indexes.

         2.3.1.3 *B-Tree Indexes:* It takes the form of a balanced tree in which every path from the root to the tree leaf is of the same length. It eliminates the redundant storage of search key values.

      2.3.2 *Hashed Index:* It is based on the values being uniformly distributed using a mathematical function called hash function.

Further details about File Organisation is dealt within Unit 2 of MLII-104.

4.5.4 Indexing Systems using AI Techniques

Artificial Intelligence (AI) is a branch of computer science concerned with study and creation of computer systems that exhibits some form of intelligence, systems that can learn new concepts and tasks, systems that reason and draw useful conclusions about the world around us, systems that can understand a natural language or perceive or comprehend a visual scene and systems that perform other types of feats that require human types of intelligence. The research in AI applications to information retrieval is gaining much importance. The recent development of the AI applications in the areas of which only those concerned with the subject indexing system is discussed below:

4.5.4.1 NLP-Based Subject Indexing System

Natural Language of Processing (NLP), one of the areas of AI, has gained momentum in research activity that explores how natural language text that is entered into a computer system can be manipulated and transformed into a form more suitable for further processing for improved storage and retrieval of information. NLP offers a potential viable alternative to statistical techniques in computerized indexing. Most automatic retrieval systems based on NLP techniques, convert the contents of the document files and user's queries in an internal form and the task of matching takes place at that level of the system, which is known as natural language interfaces, or *front-end systems*. At the core of any NLP technique there lays an important issue of natural language understanding. The process of building computer programs that understand natural language requires knowledge of how the words are formed, how the words in turn form clauses and sentences. In general, the different levels of knowledge that is to be used in natural language understanding falls into the following groups:

1) *Morphological Knowledge:* This level gives knowledge of word formation and deals with the morphological structure of words like the word root, prefix, suffix and infixes. The basic unit in a written word is a morpheme.
2) *Lexical Knowledge:* A lexicon consists of words considered as valid in the given domain. The lexicon may also contain syntactic markers or certain categories, which can be useful in processing. This level deals with thesaurus look up, spelling variations, acronyms and abbreviations, etc.
3) *Syntactic Knowledge:* Syntax deals with the structural properties and validity of input sentences—how a right combination of words in a particular sequence constitutes a valid sentence.
4) *Semantic Knowledge:* Semantics deals with the meaning of words and that of sentences. Different methods of representation of meaning have been developed over the years.
5) *Pragmatic Knowledge:* A given concept may occur in a number of different meanings, and to decide about the correct meaning in a given context the NLP system needs pragmatic knowledge. Pragmatic level deals with sentences in a particular context. This requires a higher-level knowledge, which relates to the uses of sentences in different contexts. This knowledge is useful because it helps to eliminate ambiguities and supplements the semantic representation.
6) *World Knowledge:* In order to carry out effective communication, both the communicator and communicate should have background knowledge, either to send or to receive a message, without any noise. This background knowledge is considered as the world knowledge of a particular domain.

Chomsky's classification of the types of grammar formalism may be correlated

with the above mentioned levels of knowledge. A NLP system having pragmatic and world knowledge constitutes his *'type 0'* grammars; systems with semantic knowledge or context sensitive grammars constitute his *'type 1'* grammars; systems with syntactic knowledge or *context free grammars* constitute *'type 2'* grammars; and systems with most restrictive or regular grammars constitute 'type 3' grammars. One method to formalize our knowledge is to provide a series of rewrite rules (known as grammars) that will generate the legal sentences of the language. Of the above mentioned grammars, *context free grammars* or *'type 2'* grammars, introduced by Noam Chomsky, are well understood from the computational point of view. In particular, the *Phrase structure grammars* (*PSGs*) have been used extensively in developing parsers. Another variant of it, called *Definitive Clause Grammars* (*DCGs*) is the basis for a programming language. *PROLOG (Programming in Logic)* and *LISP (List Processing)* are most popular languages in artificial intelligence programming.

For syntax analysis in NLP, valid sentences of a language are recognised and their underlying structures are determined. A central component of the syntactic phase is the *parser* (the term *parse* is derived from the Latin phrase *pars orationis* which means 'part of speech), a computational process that takes individual sentences or connected texts and converts them to some representational structure useful for further processing. *Parsing* refers to the use of syntax to determine the functions of the words in the input sentences in order to create a data structure that can be used to get at the meaning of the sentence. The major objective of the *parsing* is to transform the potentially ambiguous input phrase into an unambiguous form as an internal representation. A few of the notable computational models of the parsers are *Finite State Transition Networks (FSTN), Recursive Transition Networks (RTN), Augmented Transition Network (ATN), Definite Clause Grammar (DCG),* etc. There are many computational models of semantic grammars like *Conceptual Dependency Grammars (CDG), Modular Logic Grammars (MLG), Lexical Functional Grammars (LFG), Case Grammars (CG),* etc.

NLP has progressed to the point at which assignment indexing by computers should not be possible. Vleduts-Stokolov, for example, described an experiment performed at BIOSIS in which terms from a limited vocabulary of 600 biological concept headings were assigned automatically to journal articles. The assignment is achieved by matching article titles against a semantic vocabulary containing about 15,000 biological terms, which, in turn, were mapped to the concept headings. In NLP-based subject indexing, phrases are automatically extracted from texts to serve as content indicators. Nonsyntactic methods are based on simple text characteristics, such as word frequency and word proximity, while syntactic methods selectively extract phrases from parse trees generated by an automatic syntactic analyser.

It seems very likely that we will see increased emphasis on the use of natural language in information retrieval in future. This claim seems justified due to the following factors:

a) The continued growth in the availability of machine-readable databases many of which are in natural language form;

b) The continued expansion of online system, which is likely, eventually, to put terminal in the offices and houses of scientists and other professionals. Bibliographic searching in one of many possible applications of these terminals and natural language mode of searching seems imperative in this type of application;

c) A number of evaluation studies have indicated that natural language may offer several advantages over controlled vocabularies in many retrieval systems;

d) Natural language indexing systems have been shown to work, and to work well, in the legal field, the scientific information dissemination centres, the defence and intelligence communities; and

e) New development in computer storage devices will make the storage of very large text files increasingly feasible.

The development of the AI techniques in other two important areas concerned with the subject indexing system includes the following:

4.5.4.2 Knowledge Representation-based Subject Indexing System

Knowledge representation connotes a way of codifying knowledge. The arrangement and representation of knowledge as reflected in classification schemes and thesauri may be treated as one form of codification of knowledge. The codification helps computer to store, process and inference the codified knowledge. A computer system capable of understanding a message in natural language would seem to require both contextual knowledge and the processes for the inferences assumed by the generator. Some progress has been made towards computer systems of this sort, for understanding spoken and written fragments of language. Fundamental to the development of such systems are certain ideas about structures of contextual knowledge representation and certain techniques for making inferences from that knowledge. Based on the idea that knowledge is symbolic, the methods of using NLP tools and techniques include use of symbols and symbol structures, semantics and roles, categories and relations present in the subject statement. The work in the area of 'semantic primitives' or 'semantic factoring' leads to several means of representing knowledge for reasoning, including symbolic equations, frames and semantic nets.

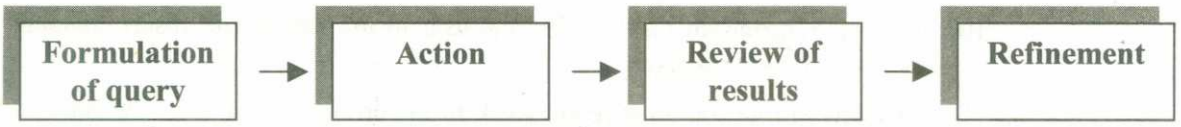
4.5.4.3 Expert System-based Subject Indexing

An *Expert system* is the embodiment, within the computer, of a knowledge-based component derived from an expert in such a form that the computer can offer intelligent advice or take an intelligent decision about the processing function. An *expert system* consists of knowledge acquisition, knowledge base system, inference machine, and user interface. This is one of the major areas of AI with a wide application to information processing and retrieval. The researches in this area lead to the development of expert systems, that should be fruitful in solving the problems of indeterminacy in both indexing and term selection for retrieval. Expert systems were designed to assist the user in query formulation and selection of relevant documents. A number of studies were undertaken in devising expert systems to aid the process of subject indexing. One notable example is the Medical Indexing Expert System or MedIndEx System (previously referred to as the Indexing Aid System) being developed at National Library of Medicine (NLM), USA. Referred to as a 'prototype interactive knowledge-based system', it uses an experimental frame language and is designed to interact with trained MEDLINE indexers by prompting them to enter MeSH terms as 'slot fillers' in completing document-specific indexing frames derived from the knowledge-base frame. Another expert system for machine-aided indexing used in the Central Abstracting and Indexing Service (CAIS) of the American Petroleum Institute (API) incorporates a rule-based expert system founded on the API thesaurus, which has been in operation since 1985. Terms automatically generated by the system (by matching abstracts against a knowledge-base derived from the API thesaurus) and by API's human indexers are reviewed by a human editor, and edited terms are added to print indexes and the online index. At present the base contains 14,000 text rules.

4.5.5 User Interface Design

User interface performs two major tasks – search or browse an information

collection and display of search results. It is also designed to perform other related tasks such as sorting, saving and/or printing of search results and modification of search query. Obviously, the success of any information retrieval system largely depends on the efficient and effective design of user interface. A well-designed user interface enhances the quality of interactions with information system and information search process my be divided into four major phases –



Shneiderman proposes following guiding principles for the design and development of user interface for computerised information retrieval –

- Strive for consistency in terminology, layout, instructions, fonts and colour
- Provide shortcuts for skilled users
- Provide appropriate and informative feedback about the sources and entity that is being searched for
- Design of message or notification system to indicate end of search process
- Permit reversal of actions so that users can undo or modify actions
- Allow users to monitor the progress of a search
- Permit users to specify the parameters to control a search
- Help users to recall their search queries
- Provide extensive online help and error-handling facilities
- Error messages should be clear and specific
- Provide facilities to enter long queries and use of Boolean, relational and positional operators
- Provide alternative interface for novice and expert users
- Provide multilingual interface if required

Information retrieval systems vary in terms of design, objectives, characteristics, contents, and users. However, the above guidelines with the help of following technical features will help to design an effective user-centred information retrieval system :

1) The presentation layer must reside within the client. Windows clients must contain an applications layer.
2) The system proposed must store all help files on the client for immediate retrieval/customisation.
3) The system proposed must enable the operator to request and receive context-sensitive help for the command in use with a single keystroke.
4) The client(s) proposed must enable the operator to page backwards and forwards through the help text, and include hypertext links to related topics, screen shots, examples, etc., and the ability to access an index of help topics.
5) The clients proposed must enable the operator to transfer any data field from one screen to another.
6) Wizards must guide the operator through a series of steps necessary to complete a defined process, without the use of commands or traditional menus.

7) Windows clients must perform all edit checking of user input before sending input to the server.

8) The clients proposed must display a form with all appropriate fields on the screen when an operator initiates a command.

9) The clients proposed must enable the user to type data onto the screen at the current cursor position.

10) The clients proposed must enable the user to use delete and insert character keys to correct mistakes.

11) The clients proposed must retain work forms on the screen, until the operator changes to another command.

12) If an error is detected, the system must report the error on the screen, leaving the form and the operator's input otherwise intact.

13) The system proposed must display an explanatory error message whenever the operator has provided inappropriate input.

14) The system proposed must set up a session with a user, which records user's search history and allows users to re-execute their previous searches against the same server or a different server or servers.

15) The system proposed must allow restriction of access to local or remote databases based upon the user's login and password.

16) The system must retain a user's authorization as he or she moves from one database to another within one session.

17) The system must allow both access by specific IP without sign-on and sign-on capability for users from outside the allowed IP range.

18) The proposed Web client must support creation and execution of simple or complex searches.

19) The proposed Web client must support browsing (SCAN) and selecting terms from standard vocabulary device.

20) The proposed Web client must support hypertext searching for related items and should support display of cross-reference information.

21) The proposed Web client must support sorting of search results to user-defined sequence.

22) The proposed Web client must support record export to printer, local file or e-mail.

23) The proposed Web client must support linking to library holdings information on multiple servers via the Z39.50 protocol.

24) The proposed Web client must support execution of previous searches from stored search history.

25) The proposed Web client must allow authentication of users by user ID and optional personal identification number, and permit authenticated users to:

   a) access their own user accounts to view the status of charges, holds, fines, and bills associated with their account; and

   b) obtain access to additional databases, such as licensed journal citation and other reference databases closed to anonymous users.

26) The system proposed must enable an authorized user to:

   a) save records to the review file one at a time, or in stated ranges;

   a) view records stored in a review file using Vendor's client software;

b) remove records from the review file;

c) duplicate records in the review file; and

d) print records in the review file;

An example of User Interface in computerised indexing is given below:

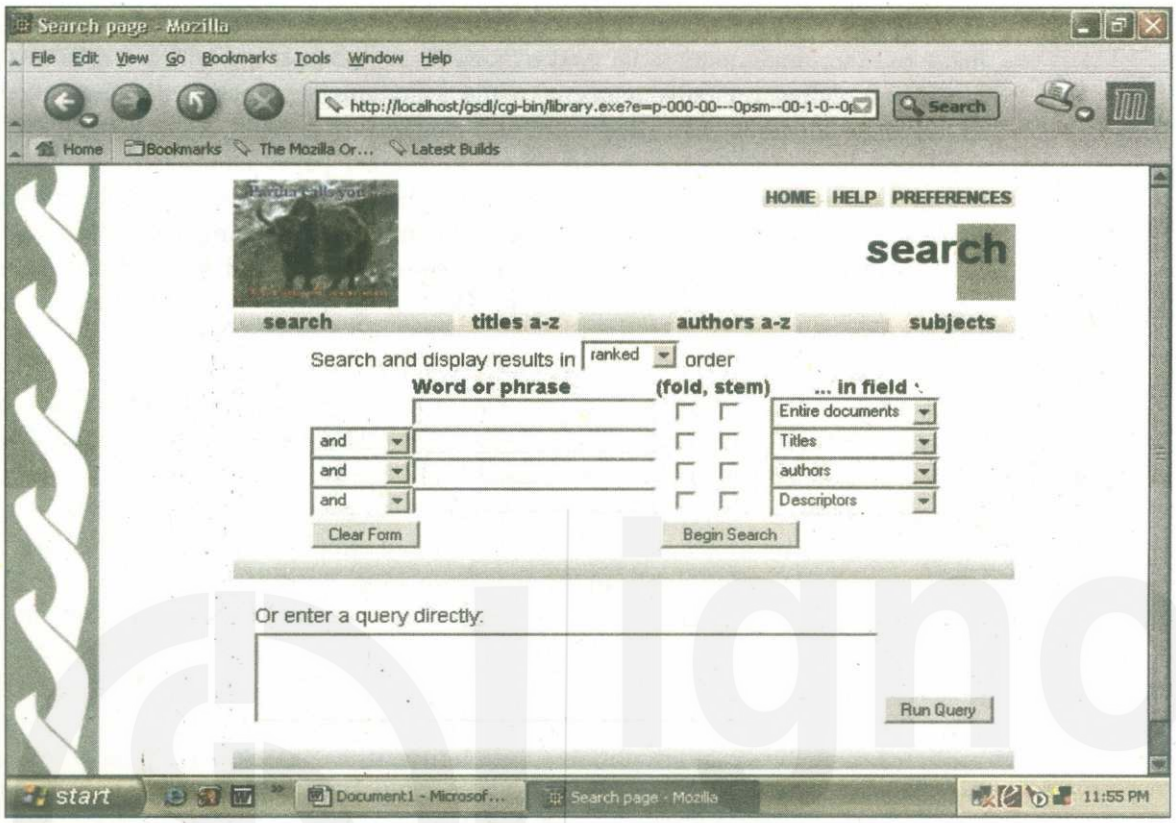


Fig. 4.7 : Screen snap shot of user interface

Self Check Exercises

10) Mention the methods adopted in measuring the word significance in computerized indexing system.

11) What are the different levels of knowledge used in natural language understanding for the NLP-based subject indexing system?

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

.....

.....

.....

.....

.....

.....

.....

4.6 NON-CONVENTIONAL INDEXING : CITATION INDEXING

Citation indexing, a variety of non-conventional indexing developed by Eugene Garfield, is the technique of bringing together the documents, which manifest association of ideas to establish the relevancy of information in a document

through mechanical sorting of citations in a citation index. A citation index, according to Garfield, is *an ordered list of cited articles each of which is accompanied by a list of citing articles.* The citing article is identified as a source, the cited article as a reference. When a document refers to some other documents it can be assumed that there is some similarity or association of ideas between the citing documents and cited documents. An author usually mentions along with his work all those documents, which he has consulted or referred to. This is called citing or referring documents. Citations are thus the references made to other documents in the text of a work. A document to which a reference is made is called a cited document, while a document, which makes reference to the cited document, is called a citing document. In general, a citation implies a relationship between a part and whole of a cited paper and a part or whole of the citing paper. The relationship between the cited documents and the citing documents forms the basis of citation indexing. By following the subsequent citations, the history of an idea can be traced—where and how it has been applied and whether it is sustained, rejected or absorbed into the latter work representing knowledge.

Though the citation indexing has provided a new approach to the bibliographic file organisation, the basic idea is not new. The first application of this idea was evident in *Shepard's Citations,* developed by Frank Shepard in 1873 as an index to American legal cases. It represents a list of American court cases, each case being followed by a complete history written in a simple code. The listing under each case displays the publications that have been referred to the cases, other court decisions that have affected the cases, and any other references that may be of value. Appreciating the usefulness of the system, Eugene Garfield stressed the need of adopting it in the field of science and technology in 1950's and in 1961 brought out an experimental Science Citation Index (SCI) after some preliminary studies. Institute for Scientific Information (ISI), Philadelphia in 1963, published the first SCI comprising the literature of 1961 and since then it is being brought out on regular basis. The online version of the SCI, known as SCISEARCH, is being published from 1974 and is available through several of the major online host systems—such as, DIALOG, etc. The compact disc version of the SCI is also available. It incorporates new feature called *related records*, which leads the searcher to other records having references in common with the ones he has already retrieved. ISI also brought out the Social Science Citation Index (SSCI) and Arts and Humanities Citation Index (A&HCI) in 1973 and 1978 respectively. The publication of the citation classics, with the first issue of *Current Contents* in 1977, forms an important and interesting venture of the ISI. Citation classics are articles selected every week by ISI from SCI, SSCI and A&HCI databases that have become classics in different fields by virtue of their high citation rate.

Science Citation Index (SCI)

SCI is the brain-child of Dr. Eugene Garfield, one time Director of the Institute for Scientific Information (ISI), Philadelphia. He created and nurtured it and brought it to the present position of the popularity, and is perhaps the best example of its kind. Initially, SCI started with two parts—*Citation Index* and *Source Index.* Since 1967, its third part, the *Permuterm Subject Index* was introduced. These three indexes, though separate are related to each other. It is published quarterly and the December issue is a cumulative issue for the whole year.

Citation Index (CI)

A CI entry contains two types of information— (a) information about a cited item (reference), and (b) information about citing items. It is arranged alphabetically by cited author, using last name of the first author (when there are more than one author). An entry for a cited item consists of the first author's

name and initials, year of publication of the cited item, details about the source document like title in abbreviated form, its volume number and starting page number. When different papers of a particular author are cited, the papers are arranged in chronological order. When different authors in different places cite a particular document, all these citing items are displayed immediately under the cited item by alphabetical order of the source authors. The information included about a citing item are: citing author's name and initials, title of the source document, its year of publication, volume number and starting page of the citing item. Coded symbols like A (for abstract), E (for editorial), N (for technical notes), etc. are used to indicate the nature of citing items.

The CI part also has two other sections on anonymous cited documents and *Patent Citation Index (PCI)*. Anonymous cited documents are arranged alphabetically by titles, and the PCI presents a list of all cited patents (foreign and domestic) in numerical sequence by patent number and usually provides the year of issuance, the inventor's name, and country.

Example (cited reference followed by citing references):

KOHARA Y				
87	CELL	50	495	
RAMAYO	GENETICS	122	65	89
WOOD	ER J BIOL CHEM	264	8297	89

Source Index (SI)

The SI is an author index of the citing items, arranged alphabetically by the last name of the first author of the source item. An entry for a citing item contains the name of the first author, the names of co-authors (up to ten), if any, the full title of the article, the title of the source document, its volume number, issue/part/ supplement number, starting page number, year, coded symbol for type of document, number of references in the bibliography of the citing item, accession number of the source periodical as filed at ISI, and author's address. The SI also has two other sections on anonymous items, and a *Corporate Index.* Anonymous items are arranged alphabetically by the titles of periodicals, and are given at the beginning of the SI. In the *Corporate Index,* all the source items are arranged alphabetically by author under the name of the organization where the work was done. When more than one organization is involved in a particular project, an entry is created for each organization.

Example:

WOOD ER

MATSON SW – THE MOLECULAR-CLONNING OF THE GENE ENCODING THE ESCHERICHIA-COLI 75-KDA HELICASE AND THE DETERMINATION OF ITS NUCLEOTIDE-SEQUENCE AND GENETIC-MAP POSITION U 5453

J BIOL CHEM 264(14) : 8297-8303 89 54R

UNIV N CAROLINA, DEPT BIOL, CHAPEL HILL, NC 27599, USA

Permuterm Subject Index (PSI)

The PSI satisfies the conventional subject approach of the users. In the PSI, significant or subject denoting terms are selected from the titles of papers appearing in the source periodicals. These terms are then permuted so that each of these terms serves as a primary term. Co-terms—that is, all other terms, which are related to a primary term, are arranged alphabetically under the given primary term like relative index of DDC.

Example (primary term followed by co-terms):

MOLECULAR-CLONNING	
GENE	COONEY A
.....	HOFFMAN E
.....	
GENETIC-MAP	WOOD ER

Search Strategy

The basic technique implies that a searcher starts with a reference or an author, he has identified through a footnote, book, encyclopedia article, subject index, or through his personal knowledge. Then, a search is to be carried out in the CI part of the SCI to identify the authors citing this reference. A second search through the SI will reveal complete bibliographical descriptions of these source items, which will enable the searcher to judge the relevancy of the documents to his query. This search may be extended in order to build a more extensive bibliography for a particular inquiry through the process, what is called 'cycling'. Once the searcher locates the citing references in the SI, he can have a look on other documents of the citing author, which may be relevant to his query. Here we see how a search involving both CI and SI operates.

When a searcher is not aware of any reference, he may start search through Permutation Subject Index. For searching a PSI, a searcher is required to compile a list of terms relevant to the search topic. Then, he should have a look on the PSI to identify one of these terms used as primary term. Simultaneously he can locate the author using the term in his title. He can then enter the SI to identify complete bibliographical description of the document. Considering co-terms along with primary term in the PSI can sharpen this search. Once this more defined subject is obtained, he locates the appropriate authors to be used in the SI. Information thus obtained can be used in the CI to locate subsequent citing sources.

Advantages

1) This index enables to link up all recent papers on a subject with papers published earlier but with which the recent ones have association of ideas.
2) While the conventional subject indexes have problems relating to terminology, this index is unaffected by problems.
3) It has already been established that this index is faster and more efficient than conventional indexes for searching.
4) It provides the best access to the interdisciplinary literature.
5) The genesis of an interdisciplinary or newly emerging subject can be traced from this index.
6) With the help of this index an author of a cited document can know how far his ideas or research results have been appreciated, applied or criticized.
7) It can be used as a tool to identify the interconnected papers on a subject to trace the history and development of it.
8) It can also be used as a measuring tool for the progress of science, as a means of automatic content analysis and as a way of evaluating journals, papers, authors, departments, and candidates for awarding prizes, appointment and promotion.
9) It is a self-updating index as each new citation of an earlier paper is automatically listed making the list of citing documents up-to-date.
10) The preparation of the citation index needs no intellectual activity since its preparation is fully computerised.

Disadvantages

1) Citation indexes stand on the foundation of the citation practice of authors. Any discrepancy at this level will generate noise at the output stage.
2) It does not provide any logical or conventional subject arrangement to which the users are accustomed. A title not in harmony with the subject content is bound to affect the citation indexing.
3) It is often complained that a user for starting the search in a citation index must know at least one reference (PSI can be used to locate a starting point).
4) A citation index retrieves only the related documents and not the contents of the documents.
5) A search in the SCI retrieves high proportion irrelevant items from the point of view of a particular searcher.

Social Science Citation Index (SSCI)

ISI brought out SSCI in 1973. It is multidisciplinary index to the literature published in the world's leading 1400 journals in social sciences, and to selected relevant items from 3,100 science journals. SSCI covers almost all the social science subjects like Anthropology, Business, Communication, Criminology, Education, Geography, History, Information Science, Library Science, Law, Linguistics, Philosophy, Sociology, etc. It appears 3 times a year and has a calendar year index. The online version of the SSCI, known as SOCIAL SCISEARCH enables speedy and accurate searches of social science literature. The compact disc version of the SSCI is available individually for the year 1986 to 1991 along with a 5-year cumulation covering 1981—1985.

The structure, format and search of the SSCI are the same as SCI. With four integrated indexes—Citation Index, Source Index, Permuterm Subject Index, and Corporate Index, one can get access to vast amount of not only the social science literature, but also to the science literature having links with the social sciences.

Arts & Humanities Citation Index (A&HCI)

ISI brought out the A & HCI in 1978 after intensive two-year marketing research. A & HCI covers more than 25 arts and humanities disciplines like Archeology, Language, Architecture, Literature, Arts, Music, Classics, Philosophy, Dance, Religious studies, Films, Television & Radio, Folklore, Theatre, History, Theology, etc. It is published three times a year.

The structure, format and search of the A & HCI are the same as those of SCI and SSCI. Like SCI and SSCI, A & HCI consists of Citation Index, Source Index, Permuterm Subject Index, and Corporate Index.

Self-Check Exercise

12) What are the different parts of Science Citation Index?

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

.....

.....

.....

.....

.....

4.7 WEB INDEXING

The Web (also known as the World Wide Web or W3) is a vast collection of files accessible to the public through the Internet, viewed through a browser, and connected by hypertext links. The information available on the Web is so vast and easily accessible; the library professionals cannot afford to ignore it. The Web consists of a very large number of things called documents, identified by names called Uniform Resource Locators (URLs). As on today the WWW contains more than three billion pages of publicly accessible information. Anyone can publish anything on the Internet. When we search the Web we can find endless examples of information that is inaccurate, inaccessible, invalid, incomplete, out of date, unreadable or simply irrelevant. Searching the Web is compared to dragging a net across the surface of an ocean. The Web was developed to link related material from different information sources. The current use of explicit links implies that the Web is static. On the contrary, the Web is highly dynamic with new information constantly becoming available (and 'old' information being removed). While an author can include links to related documents that exist at the time of writing, new documents may become available in the future. The reasons for increasing difficulties in finding the required information on the Web in time may be summed as :

- The number of data is very large.
- The mass of data is increasing rapidly.
- The data always change in structure and content.

4.7.1 Meaning and Scope

The term 'Web indexing' refers to (a) search engine indexing of the Web, (b) creation of metadata, (c) organisation of Web links by category, and (d) creation of a Website index that looks and functions like a back-of-book index. It will usually be alphabetically organised, give detailed access to information, and contain index entries with subdivisions and cross-references. In the most general sense, Web indexing means providing access points for online information materials which are available through the use of World Wide Web browsing Software. The key to this is the use of human intellectual input to analyse and categories materials, rather than reliance on computerized searching tools. It appears from the Proceedings of the Web Indexing Workshop held at the University of Melbourne during July 2000, the following issues fall within the purview of web indexing:

a) Uploading of 'traditional' indexes (and the documents to which they refer) on to the Web to provide a wider audience with access to them.

b) 'Micro' indexing of a single Web page, in order to provide users with hyper linked access points to the materials on the page.

c) 'Midi' indexing of multiple pages, largely or wholly contained within a single Web site and falling under the responsibility of a single Webmaster.

d) 'Web-wide' indexing, providing users with centralized access to widely scattered materials, which fall under a single heading (e.g. every web page dealing authoritatively with 'breast cancer').

e) 'Macro' schemes designed to simplify or unify access to large number of Web pages falling under many different headings (e.g. every web page dealing authoritatively with any medical topic).

f) The addition of comments and annotations to provide users with some guidance before they link to selected sites and pages.

The absence of librarians' skills is well reflected in the indexing approach taken by the general search engines like Alta Vista, Google, Yahoo etc. which are at the most successful in recall but definitely not in precision. In addition, the output of these search engines definitely lacks uniformity. The search tools used in the Web usually include search engines, subject directories and information or subject gateways. To put the Web indexing in proper perspective it has become imperative to familiarise with how Web works?

4.7.2 Operational Aspects of the Web

a) A Web page or home page is a document of any size—from a few lines to an entire book. A Web site might have hundreds or thousands of pages. Every Web page has the URL (Uniform Resource Locator), which acts as the address of the file accessible on the Web. e.g. *http://www.vidyasagar.ac.in/DLIS/srr/index.htm/.* Here, *http* (Hypertext Transfer Protocol) is the protocol being used to access the resource; *www.vidyasagar.ac.in* is the name of the computer where the resource is located; *DLIS/srr/index.htm* is the pathname to locate the specific file, and this portion of the URL contains information needed by the protocol to specify which document is intended.

b) By using the *http* protocol, a program (most commonly a browser but any program can do this "\web agents" and spiders are examples of such programs that are not browsers) can retrieve a document whose URL starts with "*http//:*". The document is returned to the program, along with information about how it is encoded, for example, ASCII or Unicode text, HTML, images in GIF or JPG or MPEG or PNG or some other format, an Excel or Lotus spreadsheet, etc.

c) Documents encoded in HTML (HyperText Markup Language) form can contain a combination of text, images, formatting information, and links to other documents. Thus, when a browser (or other program) gets an HTML document, it can extract the links from it, yielding URLs for other documents in the Web. If these are in HTML format, then they too can be retrieved and will yield yet more links, and so on.

d) A *spider* is a program that starts with an initial set of URLs, retrieves the corresponding documents, adds the links from these documents to the set of URLs and keeps on going. Every time it retrieves a document, it does some (hopefully useful) work in addition to just ending the embedded links.

e) One particularly interesting kind of spider constructs an *index* of the documents it has seen. This index is similar to the index at the end of a book. It has certain key words and phrases, and for each entry it lists all of the URLs that contain that word or phrase. There are many kinds of indexes and the methods of deciding what words or phrases to index and how to extract them is at the cutting edge of research. Usually, every word in the document (except, perhaps, the very common words like "and", "the", "a", and "an") is indexed.

4.7.3 Pre-requisites for Web Indexing

i) *Know the Web*

It is necessary to be familiar with the dynamic and distributed nature of the information resources on the Web, the forces shaping it, and what makes an indexable resource.

ii) *Be a User*

A web indexer needs to be a user of the Web to look for good and bad examples of indexing and to keep track of the latest technological development.

iii) *Know How to Create the Content*

It requires knowledge about the HTML. One can look at the 'Source' of the Web page to see how it is done. The ability to mock up HTML pages will be of much help for the Web indexer to know the working of the HTML.

iv) *Know about Web Site Design*

What's good / what's bad, what works / what does not, a good index is useless unless it is compatible with the browser accessing it.

v) *Know about Search Engines* and how we can add value to them.

vi) *Start looking at XML* with particular reference to what does it mean for indexing.

4.7.4 Search Engines

Web search engines (SEs) came into existence in 1994. According to "Yahoo Search Engine Directory Listing, 2003", there are over 448 major search engines. Some examples of leading /search engines include:

- Alta Vista (www.altavista.com)
- Excite (www.excite.com/search)
- Google (www.google.com)
- Hotbot (www.hotbot.com)
- Northern Light (www.northernlight.com)
- Khoj (www.khoj.com)

A SE is a searchable database of Internet files collected by a computer program (called a wanderer, crawler, robot, worm, and spider). Indexing is created from the collected files, e.g. title, full text, size, URL, etc. There are no selection criteria for the collection of files. 'Search engines allow the user to enter keywords and the search engines retrieve Web documents from its database that match the keywords entered by the searcher. The search engine doesn't wait for someone to submit information about a site. Rather, it sends out robot programs or spider (Crawler, Web crawlers) that visits publicly accessible websites following all links it comes across collecting data for search engine 'indexes'. A spider discovers new sites and updates information from sites previously visited. A spider can also be used to check links within a website.

4.7.4.1 Components of Search Engine

A search engine might well be called a search *engine service* or a *search service.* As such, it consists of three components:

1) *Spider:* Programs that traverses the Web from link to link, identifying and reading pages.

2) *Index:* Web database containing a copy of each web page gathered by the spider.

3) *Search engine mechanism:* Software that enables users to query the index and that usually returns results in relevancy ranked order.

4.7.4.2 Types of Search Engines

A search engine downloads all the information that the page contains and then examines that information to index keywords and phrases that can be used to categorise the sites. The exact method that it uses to do this, and which information

it looks at to create the index, varies according to the search engine. Those words and phrases are added to the database alongside the URL and a description of the site. Search engines can be categorised into three types on the basis of the indexing techniques employed by them:

1) *Active Search Engine:* It collects web pages information by itself. It uses a program calls "Spider" or "Web robot" to index and categorise Web pages as well as Web sites. The spider travels around the WWW in search of new sites and adds entries to their catalogue.

2) *Passive Search Engine or Subject Directories:* Search engines of this type are possibly more accurately referred to as directories. It does not seek out web pages by itself. They rely on the WWW users to submit details on their site or their favourite sites in order to build up a database. Upon receiving the submissions somebody from the search engine company trots along to the Website suggested, has a look at it, and then places the details in the right part of the database by finding the main category and sub-categories into which website should fall. For example, Yahoo directory (www.yahoo.com) has 14 main subject categories and each of these categories has many sub-categories and those sub-categories also contain their own sub-categories, and so on almost ad infinitum. Hierarchically organised directories tend to be smaller than those of the search engines, which mean that result lists tend to be smaller as well. Because subject categories are arranged by category and because they usually return links to the top level of a Website rather than to individual pages, they lend themselves best to searching for information about a subject rather than for a specific piece of information.

   Due to the size of the Web and constant transformation, keeping up with important sites in all subject areas is humanly impossible. Therefore, a guide by a subject specialist to important resources in his area of expertise is more likely than a general subject directory to produce relevant information and is usually more comprehensive than a general guide. These guides are known as *Specialized Subject Directories.* Such guides exist for virtually every topic. For examples:

   - Voice of Shuttle (http://vos.ucsb.edu) provides an excellent starting point for humanities research.
   - Film buffs should consider starting their search with the Internet Movie Database (http://us.imdb.com).

3) *Meta Search Engine:* An increasing number of search engines have led to the creation of "meta" search tools, often referred to as *multi-threaded search engines.* A meta search engine does not catalogue any web pages by itself. It simultaneously searches multiple search engines. When a query is put before this type of search engine, it forwards that query to other search engines. There are two types of meta search engines:

   a) One type searches a number of engines and does not collate the results. This means one must look through a separate list of results from each engine that was searched. It may present the same result more than once. Some engines require the searcher to visit each site to view the results, while others will fetch the results back to their own sites. When the results are brought back to the site, a certain limitation is placed on what is allowed to be retrieved. With this type of meta search engine, one can retrieve comprehensive, and sometimes overwhelming, results. An example of this type of engine is *Dogpile.*

   b) The other type is more common and returns a single list of results, often with the duplicate hits removed. This type of meta engine always brings the results back to its own site for viewing. In these cases, the engine retrieves a certain number of documents from the individual engines it has searched, cut off after a certain point as the search is processed. *Inference Find* claims to

return the maximum number of results that its targeted search engines will allow. Other meta search engines stop processing a query after a certain amount of time. Still others give the user a certain degree of control over the number of documents returned in a search. All these factors have two implications:

- These meta search engines return only a portion of the documents available to be retrieved from the individual engines they have searched; and
- Results retrieved by these engines can be highly relevant, since they are usually grabbing the first item from the relevancy-ranked list of hits returned by the individual search engines.

Some examples of meta search engines are:

a) Metacrawler (www.metacrawler.com)

b) SurfWax (www.surfwax.com)

c) Zapmeta (www.zapmeta.com)

Search engines can further be categorized by scope:

1) **General Search Engine:** It covers a range of services and facilitate Boolean search. Examples: Google, Alta Vista, etc.

2) **Regional Search Engine:** It refers to country specific search engine for locating varied resources region-wise. Examples: Euro Ferret (Europe), Excite uk (UK), etc.

3) **Subject Specific Search Engine:** It does not attempt to index the entire Web. Instead. it focuses on searching for Websites or pages within a defined subject area, geographical area or type of resource. Examples: Geo index (Geography / Environmental science), Biochemistry Easy Search Tool (Biochemistry). Because this specific search engine aims for depth of coverage within a single area, rather than breadth of coverage across subjects, they are often able to index documents that are not included even in the largest search engines databases. Some examples of subject specific search engines are:

www.123india.com	Regional
www.in.altavista.com	Regional
www.yahoo.co.uk	Regional
www.naukri.com	Employment
www.ndtv.com	News
www.zipcode.com	Weather
www.khoj.com	India-specific

4.7.4.3 Features of Search Engines

Search engines offer numerous features:

1) When using a Web search engine by entering more than one word, the space between the words has a logical meaning that directly affects the results of the search. This is known as the *default syntax*. For example: In Alta Vista, Infoseek, and Excite, a search on the words "birds migration" means that the searcher will get back documents that contain either the word "birds", the word "migration" or both. The space between the words defaults to the Boolean OR. This is probably not what the searcher wanted for this search. Again in HotBot, a search on the words "birds' migration" in Lycos and Northern Light means that the searcher will get back documents that contain both the words "birds" and "migration". Here, the space between the words defaults to the Boolean AND. Perhaps this is more appropriate.

2) Search engines return results in schematic order. Most search engines use various criteria to construct a term relevancy rating of each hit and present

the search results in this order. Criteria can include: search terms in the title, URL, first heading, HTML META tag; number of times search terms appear in the document; search terms appearing early in the document; search terms appearing close together; etc. Google page ranking algorithm displays mostly cited/hyperlinked Web sites/Web pages at the top of the screen.

3) One of the most interesting developments in search engine technology is the organization of search results by concept, site, domain, popularity and linking rather than by relevancy. Search engines that employ this alternative may be thought of as *second-generation search services*. For example:
   - Direct Hit ranks according to sites other searchers have chosen from their results to similar queries.
   - Google ranks by the number of links from pages ranked high by the service.
   - Inference Find ranks by concept and top-level domain.
   - MetaFind sorts results by keyword, alphabetically or by domain.
   - Northern Light sorts results into Custom Search Folders representing concepts and/or types of sites.
4) Often multiple pages are retrieved from a single site because all they contain the given search term. Alta Vista, Infoseek, HotBot, Northern Light and Lycos avoid this by a technique called *results grouping,* whereby all the terms from one site are clustered together into one result. It provides the opportunity to view all the retrieved pages from that chosen site. With these engines, one may get a smaller number of results from a search, but each result is coming from a different site.
5) Search engines do not index all the documents available on the Web. For example, most search engines cannot index files to password-protected sites, behind firewalls, or configured by the host server to be left alone. Still other Web pages may not be picked up if they are not linked to other pages, and are therefore missed by a search engine spider as it crawls from one page to the next. Search engines rarely contain the most recent documents posted to the Internet; do not look for yesterday's news on a search engine.
6) Contents of databases will generally not show up in a search engine result. A growing amount of valuable information on the Web is not generated from the databases. This aspect of the Web is sometimes referred to as "*the invisible Web*" because database content is "invisible" to search engine spiders.
7) Some search engines allow users to view a viewed display of the retrieved Web sites / Web pages, clustered under different topics related to the search term(s). Examples include Kartoo (http://www.kartoo.com), Vivisimo (http://www.vivisimo.com), etc.

4.7.4.4 Functions of Search Engines

There are differences in the ways various search engines work, but they all perform three basic tasks:

a) They search the Internet by using a specialized software, called crawler or robot; these software/agent can find out web pages by following hyperlinks;

b) These agent/software send the cached version of web pages to the repository of a search engine (SE) and SE keeps an index of the words they find, and where (URL) they find them; and

c) They allow users to look for words or combinations of words found in that index.

The above-mentioned functions of search engine may be illustrated through the diagram given below:

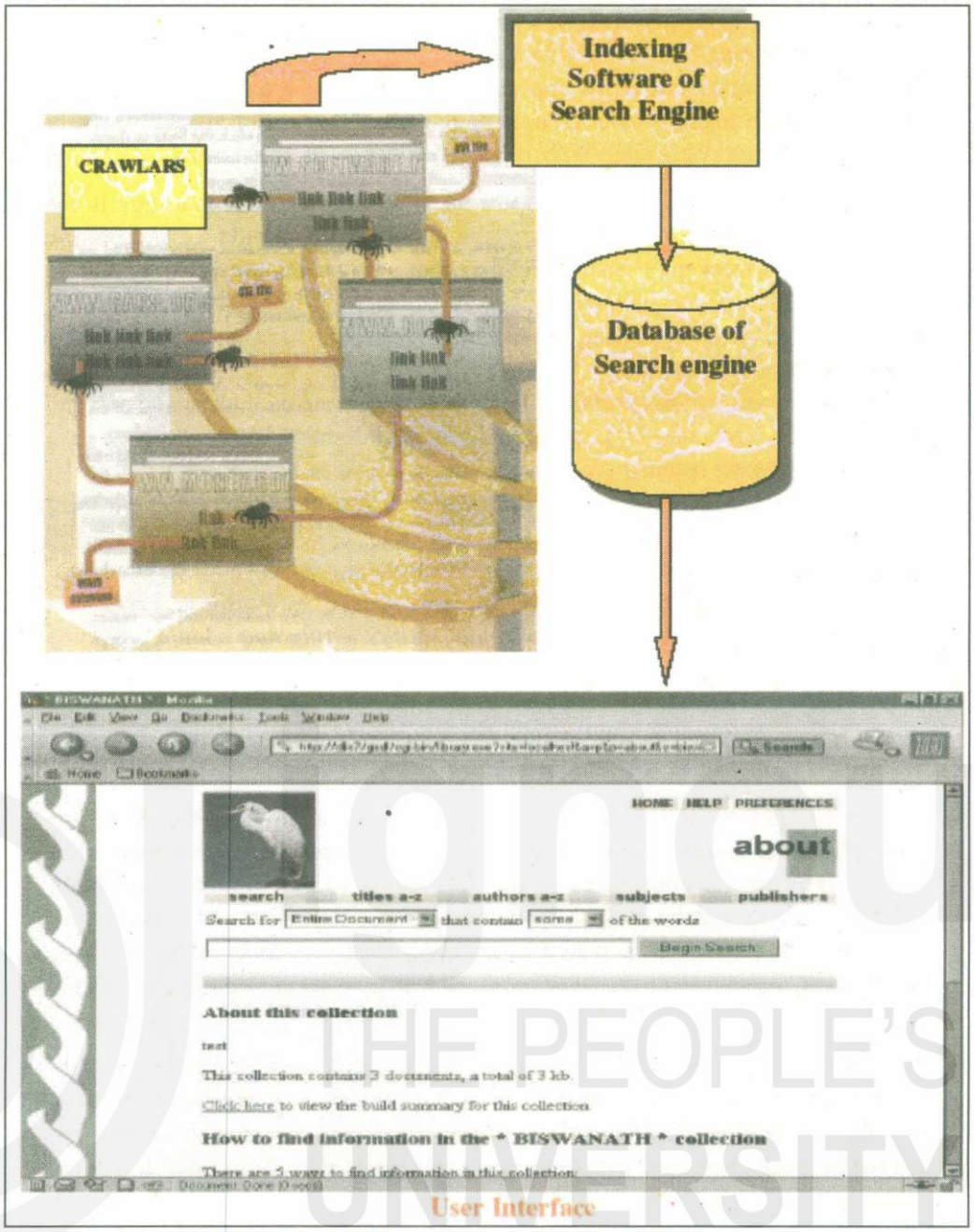


Fig. 4.8 : Diagramatic view of a search engine

4.7.4.5 Subject Directory Vs. Search Engine

We have already discussed in the section 4.7.4.2 of this Unit that a subject directory is a service that offers a collection of links to Internet resources submitted by the site creators or evaluators and organised into subject categories. Directory services use selection criteria for choosing links to include, though the selectivity varies among services. Most directories offer a search engine mechanism to query the service. When using subject directories, keep in mind that there are two basic types of subject directories: *academic and professional directories* often created and maintained by subject experts to support the needs of researchers, and *commercial portals* that cater to the general public and are competing for traffic. *INFOMINE* is a good example of an academic subject directory. *Yahoo* is a good example of a commercial portal. The points of differences between the SD and SE are shown in the following table:

Table 4.4 : Subject Directory Vs. Search Engine

Subject Directory (SD)	Search Engine (SE)
1) A SD provides categorized list of websites with brief description based on submission by Web site owners, scrutinized and edited by professional editors. It enables the searcher to move from menu to menu, making one selection after another until he gets to the level where the chosen sites are enlisted. Alta Vista, which is the largest SD today, covers more than 2 million Web sites.	1) While a SD categorizes Web sites and contains very littleinformation about them (just the description), a SE indexes all the information on all the web pages it finds. A SE deals with specific piece of information, not categories.
2) A SD takes the searcher to the home page of a Web site, and from there he can explore to get to what you want eventually.	2) A SE takes the searcher to the exact page on which thewords and phrases he is looking for.
3) With SD, a searcher requires to have a minimal knowledge to understand the catalogue.	3) With SE, one can start with one specific piece of information, like a name or phrase, and use that to find more, without knowing anything about the subject.
4) A SD is crafted by human beings, based on their judgment, like files in a file cabinet.	4) Indexes in the SE are generated automatically, based on the words and phrases that are found on Web pages. There is no human judgment filtering or rearranging of the information.
5) A SD is organised like a library.	5) Indexes in the SE are organized internally so that computer can help to find precise information from vast amount of knowledge.
6) A SD can be useful when a searcher has a vague idea of what he wants or a broad topic and wishes to view recommended sites relevant to that topic.	6) A SE is appropriate to use when a searcher is looking for a specific site or has narrow topic to pursue.
7) A SD categorises web sites and contains very little information about them (just description).	7) A SE indexes all the information on all the web pages it finds.
8) A SD requires having a minimal knowledge of the subject to understand the categories.	8) With SE, one can start with a specific piece of information, like a name or a phrase, and use that to find more, without knowing anything about the subject.

Because subject directories and search engines are so different and complement one another's capabilities, most major search engines have partnered with one or another of the major directories. For instance, Alta Vista now uses the *Open Directory Project* and the *LookSmart Directory* for its categories and Web sites. Today, the Open Directory includes over 1.5 million Web sites, in over 2, 50, 00 categories. Based on the work of over 23,000 editors, it is growing very quickly. For comparison, Yahoo now has about 1.2 million sites. Its growth is limited by what it is possible for a staff of paid editors to accomplish.

4.7.4.6 Subject Gateway

A subject gateway is recognised to be specialised in resources on a particular field and is compiled by the people, not robots. Subject Gateway or Information Gateway-type resources include Internet catalogues, subject directories, virtual libraries and gateways and these resources are organised into hierarchical subject categories. It put more focus on relevance and quality.

4.7.5 Semantic Web

The web has been said to reach its full potential only when it becomes a place where data can be shared, processed, and understood by automated systems as

well as by people. But the reality is that though the present day search engines like Google, Yahoo!, etc give us lot of hits, mostly spew out irrelevant information in answer to a search query. The problem with these search engines is that they use mostly statistical methods like frequency of occurrence of words, co-occurrence of words, etc. resulting a number of irrelevant hits against a search on the Web. Though some search engines like Google and Yahoo! use human edited entries; still they come up with a large number of wrong hits. The "semantic web" is an approach to extend the web with semantic information. Tim Berners-Lee, the developer of HTML, HTTP, URI and WWW, introduced the concept of Semantic Web. His visualisation of Semantic Web is that in future we will have intelligent software agents that will analyse a particular given situation and present us with the best possible alternatives. The idea behind Semantic Web is to develop such technologies that make the information more meaningful for the machines to process, which in turn makes search and retrieval of information more effective for searchers.

4.7.5.1 Meaning of Semantic Web

The dictionary meaning of the term "semantics" is "the study of meaning". It refers to the meaning of a string in some language, as opposed to syntax, which describes how symbols may be combined independent of their meaning. In the world of Internet the word "semantics" has slightly different meaning. In that world it means that both human and machine should have the understanding of particular topic. The World Wide Web Consortium (W3C) gives the following two definitions for the Semantic Web:

1) The Semantic Web is the representation of data on the World Wide Web. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs (Uniform Resource Identifiers) for naming.
2) The Semantic Web is the extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

The conception of Semantic Web is characterised by developing tools and technologies like languages, standards and protocols so that the Web becomes meaningful.

4.7.5.2 Key Technologies of Semantic Web

Most of the technologies involved in the development of the Semantic Web are still in their infancy. Some of them already in use are the URIs (for identifying documents uniquely and globally), XML (for structuring the data semantically), RDF (to base the structure of the documents on a common model base), Ontologies (to define the objects/entities and the interrelations between these objects/entities), etc.

1) **Uniform Resource Identifier (URI):** A URI is an identifier used to identify objects in a space. The most popularly used URI is the URL. URL is a subset of URI, they are not synonymous. A URI can be anything from the name of a person or an email address or URL or any other literal. The URI specifies a generic syntax. It consists of a generic set of schemes that identify any document/ resource like URL, URN (Uniform Resource Name), URC (Uniform Resource Characteristic), etc.
2) **XML (eXtensible Markup Language):** The emergence of HTML led to the situation where nearly everybody wanted to publish his or her own HTML-based web pages on Internet. The content problem arose rather

quickly when it became clear that there is huge amount of information on the Net, but which is mostly unreachable because of the chaos of it: The HTML didn't tell anything of the nature of the data but only the proposed layout of it. So, anybody didn't really know anything about the structure of the data of web pages. XML was designed to attach semantic to data. The HTML is about displaying information while XML is about describing information. XML is a markup language for documents containing structured information. Structured information contains both content (words, pictures, etc.) and some indication of what role that content plays (for example, content in a section heading has a different meaning from content in a footnote, which means something different than content in a figure caption or content in a database table, etc.). XML is the first step in bringing meaning to the web. While it is similar to HTML, XML allows users to add arbitrary structure to their documents by defining their own tags, but says nothing about what the structures mean. Meaning is expressed by RDF, which encodes it in sets of triples, each triple being rather like the subject, verb and object of an elementary sentence. These triples can be written using XML tags. Many have misconception that XML will replace HTML but to whatever is the case finally actual representation is done in HTML format.

3) **RDF (Resource Description Framework)**: The problem of XML is that since it characterizes the data included into elements, it does so only by allowing the schema designer to select the names of the elements 'semantically'. The RDF is thought to be the next breakthrough technology since it provides a way to build an object model from which the actual data is referenced. RDF is a foundation of processing metadata. 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. RDF metadata expresses the meaning of term and concepts in a language (XML) that is understandable to computers. Actually the RDF relies on the support of XML. RDF is said to be a framework for describing and interchanging metadata. It has been proposed as a generic framework, which is applicable to many (or any) application domains.

RDF is suitable for enhance search engines, cataloguing for describing and classifying the content available in particular web site or digital library. It can be used by intelligent software agents to facilitate knowledge sharing and exchange and in content rating in classifying resources and describing collection of pages. RDF offers :

1) A standardised syntax for writing metadata and ontologies; and
2) A standardised set of modeling primitives like instance of and subclass of relationships.

In practice RDF is a set of triplets consisting of (i)Subject, (ii) Predicate, and (iii) Object.

In RDF things are described like: "A crane is a bird" (Crane is an instance of class Bird), when "a crane" is a subject, "is" is a predicate and "a bird" is an object. A document makes assertions that particular things (people, Web pages or whatever) have properties (such as "is a sister of," "is the author of") with certain values (another person, another Web page). Subject and object are each identified by a Universal Resource Identifier (URI). The verbs are also identified by URIs, which enables anyone to define a new concept, a new verb, just as defining a URI for it somewhere on the Web. The triples of RDF form webs of information about related things. Because RDF uses URIs to encode this information in a document, the URIs ensure that concepts are not just words in

a document but are tied to a unique definition that everyone can find on the Web. Data representation in RDF is built on the triplet or *Resource, Property* and *Value:*

- **Resource (Subject):** A resource is any entity that has to be described by a URL and is equivalent to *subject* in normal English grammar. It includes the entire web pages, as well as entire elements of an XML document. An example of a resource is a draft document and its URL is *http://www.textuality.com/RDF/Why.html*
- **Property (Predicate):** A Property is any characteristics of a Resource that has a name and is equivalent to *predicate* in normal English grammar. For example: a Web page can be recognised by 'Title' or a man can be recognised by his "Name'. So both are attributes for recognition of resource 'Web page' and 'person' respectively.
- **Value (Object):** A *property* must have a *value*, which is equivalent to *object* in normal English grammar.

The concepts related the resource, property and value of the RDF model can be described through the example of RDF based metadata encoding of web resources. The example below shows the description of the homepage of Vidyasagar University by using Dublin Core Metadata Element Set (DCMES) in the framework of RDF model. It applies DCMES in the description of the property set of the homepage of the said university.

Example of RDF Data Model

Resource	Property	Value
http://www.vidyasagar.ac.in	dc:title	Vidyasagar University Home page
	dc:creator	Chakraborty, B.
	dc:subject	Academic Institute
	dc:descriptipon	Home page for the Vidyasagar University, Midnapore, West Bengal. The University established in the year 1983 under UGC Act.....
	dc:publisher	Central Library, Vidyasagar University
	dc:contributor	Sarkar, A.K.
	role=content	
	wrtiter	
	dc:date	10/12/2001
	dc:format	text/html
	dc:identifier	http://www.vidyasagar.ac.in
	dc:language	En
	dc:coverage	Education and Research
	dc:rights	Vidyasagar University

4) **Ontologies**: XML and RDF deal with metadata i.e. they deal with the description of the information available on the Web. But, if machines are expected to interact with each other or share data in the true sense of the word, then semantic interoperability is essential. For this, a formal specification

is required to explicitly define various terms and their relationships. It is here that ontologies come into picture. An ontology, as used in the context of knowledge sharing, is defined as *a specification of a representational vocabulary for a shared domain of discourse—definitions of classes, relations, functions, and other objects—is called an ontology.* In simple terms, ontology can be said to be the definition of entities and their relationship with each other. It is based on semantic net used in Artificial Intelligence. Ontologies define data models in terms of classes, subclasses and properties. For instance, we can define a man to be a subclass of human, which in turn subclass of animals that is a biped i.e. walks on two legs.

Ontologies provide taxonomies and inference rules: The taxonomy defines classes of objects and relations among them. For example, an address may be defined as a type of location, and city codes may be defined to apply only to locations and so on. Ontology may express the rule "if a city code is associated with a state code, and an address uses that city code, then that address has the associated state code. Inference rules tell the computer how to deduce and derive knowledge by combining specific types of information about specific categories.

Ontologies can be categorized into two types: general ontologies (e.g. SENSUS, Cye, WordNet, etc.) and domain-specific ontologies (e.g. GALEN—Generalized Architecture for Languages, Encyclopedias, and Nomenclatures in Medicine; UMLS—Unified Medical Language System). Ontologies are built on using XML and RDF. But, during recent past, many specific Web ontology development languages have been developed. Some notable examples are: OIL (Ontology Interface Layer) and DAML (Distributed Agent Markup Language). These two languages have now merged into a single ontology language, called DAML+OIL. In late 2001 the W3C set up a working group called *WebOnt* to define an ontology language for the Web, based on DAML+OIL.

Ontologies can enhance the functioning of the Web in many ways. They can be used in a simple fashion to improve the accuracy of Web searches—the search program can look for only those pages that refer to a precise concept instead of all the ones using ambiguous keywords. More advanced applications will use ontologies to relate the information on a page to the associated knowledge structure and inference rules.

5) **Agents**: Another crucial element of the Semantic Web is the "agent" technology. Agents are software programs that contain logical algorithms that process information and make intelligent decisions based on a pre-programmed set of rules. Agents assist users in their tasks by getting / semantic information from ontologies and annotated web pages. Inference engines perform reasoning services for agents. There are already many examples of agents in use today. Search engines, for example, perform a set of logical processes based on the search criterion we supply them. It is hoped that many different agents will be able to work together to perform even more complex tasks on the Semantic Web than are performed by agents today.

The real power of the Semantic Web will be realized when people create many programs that collect Web content from diverse sources, process the information and exchange the results with other programs. The effectiveness of such software agents will increase exponentially as more machine-readable Web content and automated services (including other agents) become available. The Semantic Web promotes this synergy: even agents that were not expressly designed to work together can transfer data among themselves when the data come with semantics.

Self Check Exercises

13) Give the meaning and scope of 'Web indexing'.

14) Discuss the functions of the different components of a Search Engine.

15) What are the key technologies involved in the development of the Semantic Web?

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

.....

.....

.....

.....

.....

.....

.....

4.8 SUMMARY

This Unit provides a comprehensive view of theoretical and practical aspects of subject indexing. It is impossible to understand the subject indexing properly without being aware of the different principles and processes associated with it. For this, different aspects of indexing principles and processes are discussed at the outset. It is followed by the discussion on the significant contributions made in the areas of pre- and post-coordinate indexing systems at different points of time since the days of Cutter. Each contribution has been discussed with illustrative examples. The traditional subject indexing systems and techniques have taken a new turn with the applications of computers in 1950s. So, computerised indexing forms an important part of this Unit. The different methods and products of computerised indexing, its differences with manual indexing, user-interface design, indexing systems using Artificial Intelligence techniques like Natural Language Processing (NLP), Knowledge Representation Model and Expert System-based subject indexing systems are discussed here. This Unit also looks at the phenomenal growth of content on the Web as an indexing problem and describes the approaches currently used to index the Web resources. In this context, different search tools used in finding the resources on the Web along with the technologies developed so far to make computers understand the semantics underlying contents of the Web resources.

4.9 ANSWERS TO SELF CHECK EXERCISES

1) The processes of subject indexing consist of two stages: (1) establishing the concepts expressed in a document; and (2) translating these concepts into the components of the given indexing language. Establishing the concepts expressed in a document calls for understanding the overall content of the document by examining the important parts of the text like title, abstract, introduction, the opening phrases of chapters and paragraphs, illustrations, tables, diagrams and conclusions. /after examining the document, the indexable concepts are to be identified and then selected in the light of the purpose for which the indexing data will be used. Finally, it is to be followed by the translation of the selected concepts into the components of the given indexing language.

2) Exhaustivity is a measure of the extent to which all the distinct topics discussed

in a document are considered for indexing. In other words, it is a measure of the number of index terms per document in an indexing system. A high level of exhaustivity increases recall. Exhaustivity is a matter of indexing policy. Specificity, on the other hand, is the degree of preciseness of the subject to express the thought content of the document. Specificity is an intrinsic quality of the indexing language itself to represent the specific subject exactly and co-extensively. A high level of specificity increases precision.

3) C. A. Cutter was the first person that first gave a generalized set of rules for subject indexing in his *Rules for Dictionary Catalog,* published in 1876. Cutter provided rules for specific as well as compound subject headings. Cutter regarded subjects as specific and classes as broad. But, in practice, it was these broad classes that Cutter entered his specific subjects. Further, according to Cutter, order of the component terms in a compound subject heading should be the one that is decidedly more significant. But Cutter could not provide any guideline as to how one will come forward to decide which one is more significant. The question of significance varies from user to user and from indexer to indexer. J. O. Kaiser started from this point where Cutter failed to provide the guideline regarding the question of significance. Kaiser, in his *Systematic Indexing*, published in 1911, prescribed that the compound subject should be analysed by determining the relative significance of the different component terms of a compound subject through classificatory approach of categorization of terms. He categorized the component terms into two fundamental categories: *concrete* and *process.* Kaiser provided the guideline that *concrete* is more significant than *process* and so, he laid a rule that a *process* should follow *concrete.*

4) J. E. L. Farradane has identified the nine relationships in his *Relational Indexing.* These nine relationships and their respective relational operators are

   a) Concurrence / 0
   b) Self-activity / *
   c) Association / ;
   d) Equivalence/ =
   e) Dimensional / +
   f) Appurtenance / (
   g) Distinctness/)
   h) Reaction / -
   i) Causation / :

5) Syntactical relationships in PRECIS are handled by means of a set of logical rules, role operators and codes. They regulate the organisation of terms in the input string by the indexer and their manipulation to generate index entries by the computer. Role operators act as instructions to the computer. Semantic relationships in PRECIS are regulated by a machine-held thesaurus that serves as a source of *See* and *See also* references in the index. A thesaurus is generated simultaneously with the preparation of input string.

6) The basic assumption leading to the development of POPSI-Specific is that subject indexing is always a specific purpose-oriented activity. But, there has always been a tradition to depend upon a designer of a subject indexing language, and such dependency always found to be inadequate to meet the specific requirement of subject indexing at the local level. Differences in requirement would call for differences in syntax of the subject proposition. It has been stated that the flexibility should be the rule of syntax, not the rigidity. Based on this assumption, POPSI tries to find out what is logically

basic, known as POPSI-Basic, and is readily amenable to the systematic manipulation to generate purpose-oriented specific versions, known as POPSI-Specific.

7) Pre-coordinate indexing involves coordination of component terms in a compound subject by the indexer at the time of indexing in anticipation of users' approach. In post-coordinate indexing, component terms in a compound subject are kept separately uncoordinated by the indexer, and the user does the coordination of terms in accordance with his requirements at the time of searching. In pre-coordinate indexing system, the rigidity of the significance order associated with the syntactical rules may not meet the approaches of all the users of the index file. But, in post-coordinate indexing, the searcher has wide options for the free manipulation of the terms at the time of searching in order to achieve whatever logical operations are required.

8) The following devices are used to eliminate false drops in post-coordinate indexing:
   a) Use of bound terms;
   b) Links;
   c) Roles; and
   d) Weighting.

9) Different versions of keyword indexing are:
   a) KWIC (Keyword –In-Context) Index;
   b) KWOC (Keyword Out-of-Context) Index;
   c) KWAC (Keyword Augmented-in-Context) Index;
   d) KWWC (Keyword-with-Context) Index;
   e) KEYTALPHA (Key Term Alphabetical) Index;
   f) WADEX (Word and Author Index);
   g) DKWTC (Double KWIC) Index;
   h) KLIC (Key-Letter-In-Context) Index;
   i) KWIT (Keyword-In-Title) Index; and
   j) SWIFT (Selected Words In Full Titles) Index.

10) Different methods adopted in measuring the word significance in computerized indexing are
   a) Weighting by location;
   b) Relative frequency weighting;
   c) Use of noun phrase;
   d) Use of thesaurus;
   e) Use of association factor; and
   f) Maximum-depth indexing.

11) Different levels of knowledge used in natural language understanding for the NLP-based subject indexing system falls into the following groups:
   a) Morphological knowledge;
   b) Lexical knowledge;
   c) Syntactic knowledge;
   d) Semantic knowledge;
   e) Pragmatic knowledge; and
   f) World knowledge.

12) There are three parts in Science Citation Index:
   a) Citation Index;
   b) Source Index; and
   c) Permuterm Subject Index.

13) The term 'Web indexing' refers to (a) search engine indexing of the Web,

(b) creation of metadata, (c) organization of Web links by category, and (d) creation of a Website index that looks and functions like a back-of-book index. It will usually be alphabetically organised, give detailed access to information, and contain index entries with subdivisions and cross-references. In the most general sense, Web indexing means providing access points for online information materials which are available through the use of World Wide Web browsing Software. The following issues also fall within the scope of Web indexing:

a) Uploading of 'traditional' indexes (and the documents to which they refer) on to the Web to provide a wider audience with access to them.

b) 'Micro' indexing of a single Web page, in order to provide users with hyper-linked access points to the materials on the page.

c) 'Midi' indexing of multiple pages, largely or wholly contained within a single Web site and falling under the responsibility of a single Webmaster.

d) 'Web-wide' indexing, providing users with centralised access to widely scattered materials, which fall under a single heading (e.g. every web page dealing authoritatively with 'breast cancer').

e) 'Macro' schemes designed to simplify or unify access to large number of Web pages falling under many different headings (e.g. every web page dealing authoritatively with any medical topic).

f) The addition of comments and annotations to provide users with some guidance before they link to selected sites and pages.

14) Functions of the different components of a search engine are :

a) **Spider**: Computer program that visit websites following all links it comes across collecting data for search engine indexes, identifying and reading Web pages;

b) **Index**: A searchable database containing indexing terms created from the Web pages by the spider; and

c) **Search engine mechanism**: Software that enables users to query the index and that usually returns results in relevancy ranked order.

15) Key technologies involved in the development of the Semantic Web are :

a) Uniform Resource Identifier (URI);

b) eXtensible Markup Language (XML);

c) Resource Description Framework (RDF);

d) Ontology; and

e) Agent software.

4.10 KEYWORDS

Amplified Phrase Order : The order of component terms in a phrase achieved by using the necessary prepositions in between them. It is a corollary of the order of significance provided by E J Coates.

Analet : Two or more isolates linked by relational operators according to J E L Farradane's Relational Indexing system constitute *Analet.*

Analysis : It refers to the conceptual analysis, which involves deciding what a document is about—that is, identification of different component ideas (concepts) associated with the thought content of the document.

Associative Classification: It refers to the association of a subject with other subjects without reference to its COSSCO relationships and results in a relative index.

Chain Indexing : A method of deriving alphabetical subject entries from the chain of successive subdivisions leading from the general to most specific level needed to be indexed. For this, it takes the class number of the document concerned from a preferred classification scheme for deriving subject index entries.

Citation Index : An ordered list of cited articles (references), each of which is accompanied by a list of citing articles (sources).

Citation Indexing : Techniques of bringing together the documents (cited documents) which manifest association of ideas to establish the relevancy of information in a document (citing document) through mechanical sorting of citations in a citation index. The relationship existing between the cited documents and citing documents forms the basis of *Citation indexing*.

Concrete : One of the fundamental categories propounded by Kaiser, which refers to things, place and abstract terms not signifying an action.

Consistency in Indexing : It is a measure that relates to the work of two or more indexers.

Controlled Vocabulary : A controlled vocabulary refers to an authority list of terms showing their interrelationships and indicating ways in which they may usefully be combined to represent specific subject of a document.

COSSCO Relationships : It refers to Coordinate-Superordinate-Subordinate-Collateral relationships in organising classification.

Default Syntax : A logical meaning of the use of space between the words when entering more than one word in a Web search engine carries out a search.

Exhaustivity : The use of enough terms to cover the all topics discussed in a document. It relates to the breath of coverage in indexing. It is sometimes called *depth indexing*.

Expert System : It is the embodiment, within the computer, of a knowledge-based component derived from an expert in such a form that the computer can offer intelligent advice or take an intelligent decision about the processing function.

eXtensible Markup Language (XML) : A subset of Standard Generalized Markup Language (SGML), a widely used international text-processing standard. XML is being designed to bring the power and flexibility of generic SGML to the Web, while maintaining interoperability with full SGML and HTML. XML is the first step in bringing meaning to the Web.

False Drops : Retrieval of unwanted items because of the false coordination of terms at the time of searching.

Hypertext Markup Language (HTML) : The standard text-formatting language for documents on the World Wide Web. HTML text files contain content that is rendered on a computer screen and markup, or tags that can be used to tell the computer how to format that content. HTML tags can also be used to encode metadata and to tell the computer how to respond to certain user actions, such as a mouse click.

Links : Special symbols used to group all the related concepts in a document separately for the elimination of false drops.

Index : A systematic guide to the contents of documents, comprising a series of entries, with headings arranged in alphabetical or other chosen order and with references to show where each item indexed is located

Indexing : The process of evaluating information entities and creating indexing terms, normally subject or topical terms, that aid in finding and accessing the entity. Index terms may be in natural language or controlled vocabulary or a classification notation.

Indexing Language : An indexing language is an artificial language consisting of a set of terms and devices for handling the relationship between them for providing index description. It is also referred to as a *retrieval language.*

Indexing Program : Computer software used to order things; frequently used to refer to software that alphabetizes some or all of the terms in one or more electronic documents.

Input String : It refers to a set of terms arranged according to the role operators in PRECIS.

Invisible Web : It refers to those pages what we can't see in the results pages after we run a search on the Web. Here, the pages remain hidden due to a variety of technical reasons, to search engines excluded them because the pages do not allow them access and also to pages, which restrict entry through the log-in password method.

Item Entry System : A type of post coordinate indexing system in which items are posted on the term. A single entry is prepared for each item, which permits access to the entry from all appropriate headings.

Keyword Indexing : An indexing system based on the usage of natural language terminology for deriving the index entries. Significant words denoting the subject, known as keywords, are taken mainly from the title and/or sometimes from abstract or full text of the document for the purpose of indexing.

Knowledge Representation : Method of codifying knowledge to enable a computer to store, process and to draw inference from the codified knowledge.

Metadata : In general, it is data about data. Functionally, a metadata is structured data about data, which describes the characteristics of a resource. A metadata record consists of a number of pre-defined elements representing specific attributes of a resource, and each element can have one or more value.

Meta Search Engine : A search engine that simultaneously searches multiple search engines in response to a query.

Meta tag : The HTML element used to demarcate metadata on a Web page.

Natural Language Processing (NLP) : It refers to the area that attempts to make the computer understand natural language.

Ontology : A specification of a representational vocabulary for a shared domain of discourse—definitions of classes, relations, functions, and other objects.

Organising Classification : It refers to the categorization of concepts and their organsation on the basis of genus-species, whole-part, and other inter-facet relationships. It is used to distinguish and rank each subject from all other subjects with reference to its COSSCO relationships.

Parser : A computational process that takes individual sentences or concerned texts and converts them to some representational structure useful for further processing.

Parsing : It refers to the use syntax to determine the functions of words in the input sentences in order to create a data structure that can be used to get at the meaning of the sentence.

Post Coordinate Indexing : Component terms in a compound subject are kept separately uncoordinated by the indexer and the searcher does the coordination of the component terms at the time of searching. Also called *Coordinate indexing.*

Pre-coordinate Indexing : the indexer carries out Coordination of component terms in a compound subject at the time of indexing by following the syntactical rules of a given indexing language.

Process : One of the fundamental categories propounded by Kaiser, which includes mode of treatment of the subject by the author, an action or process described in a document, and an adjective related to the *Concrete* as component of the subject.

Quality of Indexing : The ability to retrieve what is wanted and to avoid what is not wanted.

Relational Indexing : An indexing system developed by J E L Farradane, which involves the identification of the relationship between each pair of terms of a given subject

statement and representation of those relations by relational operators.

Relational Operators : Special symbols used to link the isolates in *Relational Indexing* to create *analets.*

Relative Index : An index showing various aspects of an idea and its relationship with other ideas.

Resource Description Framework (RDF) : A generic framework for describing and interchanging metadata on the Internet. RDF metadata expresses the meaning of terms and concepts in the XML that is understandable to computers.

Role Operators : These refer to a set of notations, which specifies the grammatical role or function of the term which follows the operators and which regulates the order of the terms in an input string in PRECIS. The rules associated with role operators serve as computer instructions for generating index entries, determines the format, typography and punctuation associated with each index entry.

Role : A symbol attached to the index term to indicate the context in which the term has been used.

Search Engine : A searchable database of Internet files collected by a computer program (called a *wanderer, crawler, robot, worm,* and *spider.*

Semantic Net : A directed graph in which nodes represent entities and arcs entities. Arcs are labelled with the names of the relation types—that is, the binary relationship to which the relationship belongs. A single node represents a single entity. Semantic nets can be used to represent various types of knowledge.

Semantics : Semantics is a study of meaning. In an indexing language, semantic relationship refers to the hierarchical and non-hierarchical relationships between the subjects and is governed by *see* and *sees also* references in an index file. Controlled vocabulary serves as the source for *see* and *see also* references.

Semantic Web : The term *Semantic Web,* introduced by Tim Berners-Lee, is the extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The infrastructure of the Semantic Web would allow machines as well as humans to make deductions and organize information. The architectural components include semantics (meaning of the elements), structure (organization of the elements), and syntax (communication).

Software Agent : A computer program that carries out tasks on behalf of another entity. Frequently used to reference a program that searches the Internet for information meeting the specified requirements of an individual user.

Specificity : It refers to the use of much smaller number of terms to cover only the central subject matter of a document. The more specific the terms used, the fewer the entries per term on the average. Specificity is the property of the vocabulary used in indexing and it relates to the depth of treatment of the content of a document in indexing.

Spider : A computer program used in indexing and retrieving Web resources with reference to their URLs that contain the given keywords or phrases. A *spider* traverses the Web from link to link, identifying and reading pages. Also called *crawler, robot,* etc.

Standard Generalized Markup Language (SGML) : A non-proprietary language/enabling technology for describing information. Information in SGML is structured like a database, supporting rendering in and conversion between different formats. Both XML and later versions of HTML are instances of SGML.

Subject Directory : A search engine service that offers a collection of links to Internet resources submitted by the site creators or evaluators and organised into subject categories.

Subject Gateway : A subject gateway is recognised to be specialised resources on a particular field and is compiled by the people, not robots. The resources in the subject gateway include Internet catalogues, subject directories, virtual libraries and gateways and these resources are organised into hierarchical subject categories. Also called *Information Gateway.*

Syntax : The grammatical structure consisting of a set of rules that govern the sequence of occurrence the terms in a subject heading.

Term Entry System : A type of post coordinate indexing system in which index entries for a document are made under each of the component terms associated with the thought content of the document. Here, terms are posted on the item.

Term Relationship : It refers to the relationship between two or more equally concrete things or phrases in a subject, which may lead to the absence of order of significance and modification of the amplified phrase order as propounded by E J Coates in his subject indexing system. Coates has identified 20 different kinds of relationships by means of prepositions: *of, for, against, with,* and *by.*

Term Significance : It refers to the order of significance developed by E J Coates in his subject indexing system. It states that the most significant term in a compound subject heading is the one that is most readily available in the memory of the enquirer and this leads to the order of significance as *Thing-Part-Material-Action.*

Uniform Resource Identifier (URI) : The syntax for all names/addresses that refer to resources on the Web.

Uniform Resource Locator (URL) : A technique for indicating the name and location of Internet resources. The URL specifies the name and type of the resource, as well as the computer, device and directory where the resource may be found. The URL is a subset of URI. For example, the URL for Dublin Core Metadata Initiative is http://dublincore.org/.

Uniform Resource Name (URN) : A URI (name and address of an object on the Internet) that has some assurance of persistence beyond that normally associated with an Internet domain or host name.

Web Indexing : Web indexing means providing access points for online information materials, which are available through the use of World Wide Web browsing Software. .

Weighting : A method of the allocation of values to indexing terms by using quantitative figures according to their importance in the document.

World Wide Web (WWW) : The panoply of Internet resources (text, graphics, audio, video, etc.) that is accessible via a Web browser. Also called Web or W3.

4.11 REFERENCES AND FURTHER READING

Borko, Harold and Bernier, Charles L. (1978). *Indexing concepts and methods*. New York: Academic Press, 1978.

Chakraborty, A.R. and Chakraborty, Bhubaneswar. (1984). *Indexing: principles, processes and products*. Calcutta: World Press.

Chowdhury, G.G. (2004). *Introduction to modern information retrieval*. 2nd ed. London: Facet Publishing.

Ghosh, S.B. and Satpathi, J.N. (eds.) (1998). *Subject indexing systems: concepts, methods and techniques*. Calcutta: IASLIC.

Gopinath, M.A. (1999). *Indexing process and models*. In: MLIS-03, Block 2 course materials. New Delhi: Indira Gandhi National Open University.

Guha, B. (1983). *Documentation and information: services, techniques and systems*. 2nd rev ed. Calcutta: World Press.

Lancaster, F.W. (1979). *Information retrieval systems: characteristics, testing, and evaluation*. 2nd Ed. New York: John Wiley.

Lancaster, F.W. (1998). *Indexing and abstracting in theory and practice*. 2nd ed. Champaign Illinois: University of Illinois.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4), 309-317.

Prasad, A.R.D. (ed.) (1993). *Artificial Intelligence applications to library and information work*. DRTC Refresher Workshop (May 26-28, 1993). Bangalore: Documentation Research and Training Centre, Indian Statistical Institute.

Prasad, A.R.D. (ed.) (1993). *Semantic web*. DRTC Workshop (December 8-10,

2003). Bangalore: Documentation Research and Training Centre, Indian Statistical Institute.

Quinn, B. (1994). Recent theoretical approaches in classification and indexing. *Knowledge Organization*, 21(3), 140-147.

Ranganathan, S.R. (1964). Subject heading and facet analysis. *Journal of Documentation*, 20, 101-119.

Salton, G. and McGill, Michael I. (1983). *Introduction to information retrieval*. New York: McGraw-Hill.

Sarkhel, J.K. and Mukhopadhyay, P.S. (2002). *Towards a semantic web: a metadata approach*. In: *National Seminar on Information Management in Electronic Libraries* (February 26-27, 2002). Kharagpur: Indian Institute of Technology Kharagpur.

Sarkhel, J.K. (2001). *Information analysis in theory and practice*. Kolkata, Classique Books.

Vickery, B.C. (1970). *Techniques of information retrieval*. London: Butterworth.

ignou
THE PEOPLE'S
UNIVERSITY

UNIT 5 EVALUATION OF INDEXING SYSTEMS

Structure

5.0 Objectives

5.1 Introduction

5.2 Purpose of Evaluation

5.3 Levels of Evaluation

- 5.3.1 System Effectiveness
- 5.3.2 Cost Effectiveness
- 5.3.3 Cost-benefit Evaluation

5.4 Evaluation Criteria

- 5.4.1 Recall and Precision
- 5.4.2 Other Performance Measures
- 5.4.3 Relevance

5.5 Evaluation Methodology

5.6 Evaluation Experiments

- 5.6.1 The Cranfield Tests
- 5.6.2 MEDLARS Test
- 5.6.3 SMART Retrieval Experiment
- 5.6.4 TREC Experiment

5.7 Summary

5.8 Answers to Self Check Exercises

5.9 Keywords

5.10 References and Further Reading

5.0 OBJECTIVES

In the previous Units, you have learnt the fundamentals of information processing and organisation. You have also been acquainted with different types of indexing systems developed for organisation of bibliographic information in the context of development of information storage and retrieval systems. Different systems have originated to suit different types of requirements. For having an efficient system to suit particular needs it is necessary the system developed is evaluated to judge its efficiency in retrieval. The present Unit discusses various aspects of evaluation. Indexing is a sub-system of the overall ISAR system. Though the Unit is discussed in the context of IR as a whole system, the methodology is equally applicable in the context of indexing systems.

After reading this Unit, you will be able to:

- analyse the reasons for evaluation of IR systems;
- understand different levels and criteria for evaluating an IR system;
- learn the methodology for evaluating an IR system;
- develop strategies in evaluating an IR system; and
- know the different evaluation experiments.

ignou THE PEOPLE'S UNIVERSITY

5.1 INTRODUCTION

Evaluation of an information retrieval system essentially means measuring the performance of the system, success or failure, in terms of its retrieval efficiency (ease of approach, speed and accuracy), and its internal operating efficiency, cost effectiveness and cost benefit to the managers of the system. In other words, we evaluate a system in order to ascertain the level of its performance or its value. An indexing system is a sub-system of an information retrieval system and hence, its performance is directly linked with its overall performance of the entire information retrieval system. An information retrieval system can be evaluated by considering the following two issues:

i) how well efficiently the system is satisfying its objectives, that is, how well it is satisfying the demands placed upon it, and

ii) whether the system justifies its existence.

We evaluate a system in order to ascertain the level of it's performance or value in terms of its *effectiveness* and *efficiency*. By effectiveness we mean the level up to which the given system attains its stated objectives. In an information retrieval system, the effectiveness may be a measure of how far it can retrieve relevant information withholding non-relevant information. By efficiency we mean how economically the system is achieving its objective. In an information retrieval system efficiency can be measured by such factors, such as at what minimum cost and effort does the system function effectively. It may be necessary that the cost factors are to be calculated indirectly, such as response time (i.e. that is time taken by the system to retrieve the information), user effort (i.e. the amount of time and effort required by a user to interact with the system and analyse the output retrieved in order to get the required information), the cost involved, and so on. Evaluation is generally done for computer-based systems. It can also be possible in case of manual system (Card catalogues) and other types of bibliographical products and services, including printed indexes.

5.2 PURPOSE OF EVALUATION

As it has been mentioned earlier, the purpose of evaluation in general is to judge the efficiency of a system in relation to retrieval, identify the shortcomings, if any, rectify them and improve upon the system.

An IR system is evaluated to obtain two types of data:

a) Performance figures for a representative group of searches; and

b) Examples of system failures to allow analysis of causes of failures in searches.

Evaluation studies investigate the degree to which the stated goals or expectations have been achieved or the degree to which these can be achieved. The above-mentioned data, when analysed, and interpreted, should yield recommendations from which decisions can be made on how best the system performance may be improved. Evaluation studies have one or more of the following purposes:

- To show at what level of performance the system is now operating;
- To compare the performance of two or more systems against a standard or norm;
- To determine whether and how well goals or performance expectations are being fulfilled;
- To identify the possible sources of system failure or inefficiency with a view to raising the level of performance at some future date;

- To justify the system's existence by analysing the costs and benefits;
- To explore techniques for increasing performance effectiveness;
- To establish a foundation of further research on the reasons for the relative success of alternative techniques; and
- To improve the means employed for attaining the objectives or to redefine the goals in view of research findings.

5.3 LEVELS OF EVALUATION

An IR system may be evaluated at various levels. F.W. Lancaster has identified the following levels of evaluations:

5.3.1 System Effectiveness

It is the evaluation of the system performance in terms of the degree to which it meets the users' requirements. It considers the users' satisfaction level and is measured by determining the utility of information to the users in response to a user query. Precision has been widely used to measure the retrieval effectiveness. It takes into consideration cost, time and quality criteria.

1) Cost Criteria:
   i) Monetary cost of user (i.e. cost incurred per search, per subscription, per document);
   ii) Other less tangible cost considerations—such as, Users' effort involved :
      - in learning the working of the system;
      - in actual use;
      - in getting the documents through back-up document delivery systems; and
      - in retrieving information from the retrieved documents.
2) Time Criteria:
   i) Time taken from submission of query to the retrieval of bibliographical references;
   ii) Time elapsing from submission of query to the retrieval of documents and the actual information; and
   iii) Other time considerations—such as, waiting time to use the system such as online.
3) Quality Criteria :
   i) Coverage of database;
   ii) Completeness of output (Recall);
   iii) Relevance of output (Precision);
   iv) Novelty of output; and
   v) Completeness and accuracy of data.

5.3.2 Cost Effectiveness

There may be many systems where system effectiveness may be there but cost effectiveness may not be satisfactory. A cost effectiveness evaluation relates measures of effectiveness to measure of cost. It is the evaluation in terms of how to satisfy user requirements in the most efficient and economical way. It takes into considerations:

a) Unit cost per relevant citation retrieval;

b) Unit cost per new, that is, previously unknown, relevant citation retrieval; and

c) Unit cost per relevant document retrieval.

5.3.3 Cost-benefit Evaluation

It is the evaluation to assess the worthiness of the system. A cost-benefit study attempts to relate the cost of providing some service to the benefits of having the service available. A number of studies have been conducted so far to determine the costs of information retrieval systems or subsystems.

Self Check Exercises

1) How do you distinguish between the efficiency and effectiveness of an IR system?

2) What are the different levels of evaluation?

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

.....

.....

.....

.....

.....

5.4 EVALUATION CRITERIA

During 1950s Perry and Kent while bring out the concept of evaluation for IR systems suggested the following:

i) **Resolution factor**: The proportion of total items retrieved over a total number of items in the collection.

ii) **Pertinency factor**: The proportion of relevant items retrieved over a total number of retrieved items. This factor is popularly named as the *precision ratio* in the subsequent evaluation studies.

iii) **Recall factor**: The proportion of relevant items retrieved over a total number of relevant items in the collection.

iv) **Elimination factor**: The proportion of non-retrieved items (both relevant and non-relevant) over the total items in the collection.

v) **Noise factor**: The proportion of retrieved items which are not relevant. This factor is considered as the complement of the *pertinency factor.*

vi) **Omission factor**: The proportion of non-relevant items retrieved over the total number of non-retrieved items in the collection.

Perry and Kent suggested the following formulae for the estimation of the above mentioned evaluation criteria:

L / N =	Resolution factor	(N—L) / N =	Elimination factor
R / L =	Pertinency factor	(L—R) / L =	Noise factor
R / C =	Recall factor	(C—R) / C =	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

C.W. Cleverdon [1962] identified six criteria for the evaluation of an information retrieval system. These are:

i) **Recall**: It refers to the ability of the system to retrieve all the relevant items;

ii) **Precision**: It refers to the ability of the system to retrieve only those items that are relevant;

iii) **Time lag**: It refers to the time gap between the submission of a request by the user and his receipt of the search results.

iv) **User Effort**: It refers to the intellectual as well as physical effort required from the user in obtaining answers to the search requests. The effort is measured by the amount of time user spends in conducting the search or negotiating his enquiry with the system. Sometimes, response time may be good, but user effort may be poor.

v) **Form of presentation** of the search output, which affects the user's ability to make use of the retrieved items, and

vi) **Coverage of the collection**: It refers to the extent to which the system includes relevant matter. It is a measure of the completeness of the collection.

Vickery identifies six criteria under two sets as follows:

Set 1

i) *Coverage*: the proportion of the total potentially useful literature that has been analysed,

ii) *Recall*: the proportion of such references that are retrieved in a search, and

iii) *Response time*: the average time needed to obtain a response from the system.

These three criteria are related to the availability of information, while the following three are related to the selectivity of output:

Set 2

i) *Precision*: the ability of the system to screen out irrelevant references,

ii) *Usability*: the value of the references retrieved, in terms of such factors as their reliability, comprehensibility, currency, etc., and

iii) *Presentation*: the form in which search results are presented to the user.

All these factors are related to the system parameters, and thus in order to identify the role played by each of the performance criteria mentioned above, each must be tagged with one or more system parameters. Salton and McGill identified the various parameters of an information retrieval system as related to each of five evaluation criteria. Table 5.1 presents the relationships between these evaluation criteria and the system parameters as identified by Salton and McGill.

Table 5.1 : Relationship between Evaluation Criteria and System Parameters

Evaluation Criteria	System Parameters
Recall and precision	1) Indexing exhaustivity Recall tends to increase the exhaustivity of indexing terms. 2) Term specificity Precision increases with the specificity of the index terms. 3) Indexing language Availability of measures for recognition of synonyms, term relations, etc., which improve recall. 4) Query formulation Ability to formulate an accurate search request. 5) Search strategy Ability of the user or intermediary to formulate an adequate search strategy.
Response time	1) Organisation of stored documents. 2) Type of query. 3) Location of information centre. 4) Frequency of receiving users' queries. 5) Size of the collection.
User effort	1) Accessibility of the system. 2) Availability of guidance by system personnel. 3) Volume of retrieved items. 4) Facilities for interaction with the system.
Form of presentation	1) Type of display device. 2) Nature of output – bibliographic reference, abstract, or full text.
Collection coverage	1) Type of input device and type of size of storage device. 2) Depth of subject analysis. 3) Nature of users' demands. 4) Nature of core subject area. 5) Physical forms of documents.

Some of the performance criteria mentioned above can be measured easily. For example, the parameters related to the collection coverage, and forms of presentation are related to policy matters, and thus are defined by the system managers beforehand. Response time and user effort can be measured without much difficulty. However, the two other criteria, recall and precision, cannot be measured so easily. In fact, measurements of these factors often cause a number of problems for the system investigators. Much research effort in the area of the evaluation of information retrieval system has concentrated on these two factors.

5.4.1 Recall and Precision

The most important parameters used for evaluating an indexing system are: *Recall* and *Precision.* The term *recall* refers to a measure of whether or not a particular item is retrieved or the extent to which the retrieval of wanted items occur. *Recall ratio* is nothing but the proportion of relevant items retrieved and thus, it is a measure of the completeness of a search in an index file. The *recall ratio* has been variously called as *hit rate, sensitivity*, and *conditional probability of a hit.* The general formula for calculation of *recall ratio* may be stated as:

$$\text{Recall Ratio} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant in the collection}} \times 100$$

Example: If a database holds 50 relevant documents as answer to a query and retrieves only 35 of them against the enquiry as output, the *recall ratio* is (35 / 50) × 100 = 70%.

The term *precision* relates to the ability of an indexing system not to retrieve irrelevant items. *Precision ratio* is nothing but the proportion of retrieved items that are relevant. By it, we measure how precisely an indexing system functions. It is quite obvious that when the system retrieves items that are relevant to a given query it also retrieves some documents that are not relevant. These non-relevant items affect the success of the system because the user, for whom it results in the wastage of significant amount of time, must discard them. The *precision ratio* is sometimes referred to as a *relevance ratio.* The general formula for calculation of precision ratio may be stated as:

$$\text{Precision Ratio} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of items retrieved}} \times 100$$

Example: If a system retrieves 50 documents as answer to a query as output and only 8 of them are relevant to that particular query, the *precision ratio* is (8 / 50) × 100 = 16%.

In the context of the evaluation, the following matrix as a common frame of reference would be useful:

	User relevance decision		
Retrieved	a Hits	b Noise	c Total retrieved
Not retrieved	c Misses	d Correctly rejected	c+d Total not retrieved
Total	a + c Total relevant	b + d Total not relevant	a + b + c + d Total collection

It appears from the above noted matrix that the system retrieves *'a'* relevant documents (i.e. *'hit'*) along with *'b'* non-relevant documents (i.e. 'noise'), and out of the remaining [*c+d*] documents, the system misses *'c'* documents that should have been retrieved, but it correctly rejects *'d'* documents that are not relevant to the given query. From this, the recall and precision ratios can be calculated as:

$$\text{Recall ratio} = [a / (a+c)] \times 100$$

$$\text{Precision ratio} = [a / (a+b)] \times 100$$

During the evaluation programme, values (a + b) can be established easily, but it is difficult to determine values (c + d). This can be done by asking the enquirer to verify all the non-retrieved documents (c + d) and judge which of them are relevant (c) and which are not (d).

An ideal indexing system attempts to achieve 100% recall and 100% precision, i.e. it attempts to retrieve all the relevant documents only. However, this is not possible in practice because as the level of recall increases, precision tends to decrease. Lancaster therefore states that *recall and precision tend to vary inversely in searching*. By this we mean that when we broaden a search to

achieve better recall, precision tends to go down. Conversely, when we restrict the scope of a search by searching more stringently in order to improve the precision, recall tends to deteriorate. In fact, by conducting a search or whole group of searches at varying strategy levels, from very broad to very specific, a series of performance points can be obtained in terms of recall and precision. If recall and precision ratios are derived for each of these search approaches and if these ratios were plotted against each other, the plot would result in *performance curve,* as shown below:

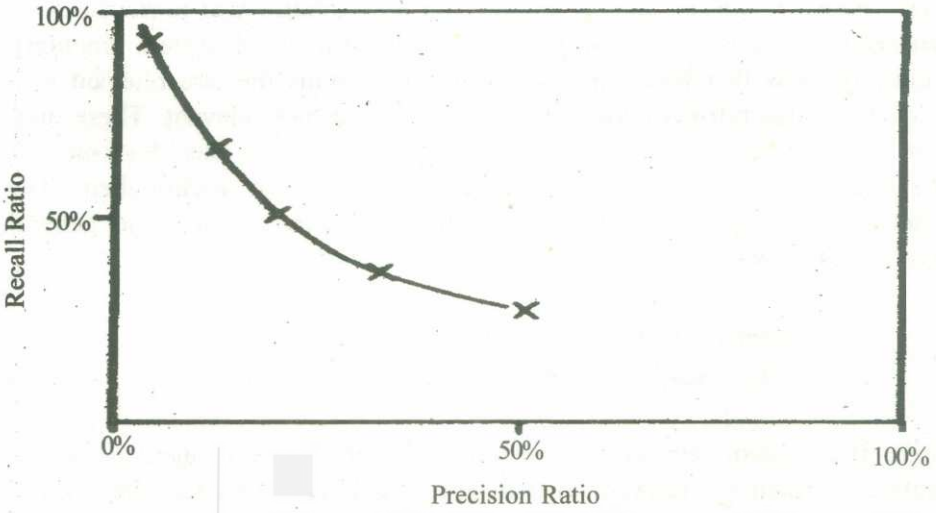


Table 5.1 : Performance curve

Fugmann has questioned the view of Lancaster in respect of recall—precision inverse relationship and by several examples he has shown that:

a) an increase in precision is by no means always accompanied by a corresponding decrease in recall, and

b) an increase in recall is by no means observed to have always in its wake a decrease in precision.

5.4.1.1 Factors Affecting Recall and Precision

Two performance criteria, i.e. recall and precision are influenced by the following factors:

a) Indexing policy: exhaustivity and specificity;

b) Requests that imperfectly represent information needs; and

c) Search strategy.

The effectiveness of an indexing system is governed by the indexing policy—*exhaustivity* and *specificity* (See the Section 4.2.7 under the Unit 4 of this course), which affect two most important parameters—*Recall* and *Precision* used for evaluating an indexing system.

A high level of *exhaustivity* of indexing tends to ensure high *recall.* For example, if we do not recognise the concept E in indexing a document dealing with five concepts A, B, C, D, E, we will never be able to retrieve it in response to a request for literature on E unless the concept E is related, hierarchically or non-hierarchically, to the concepts A to D. A high level of *exhaustivity* of indexing also tends to reduce *precision.* If we recognise all or substantial proportion of the indexable concepts for every document input, we will tend to index many concepts that are treated in only a very minor way in the documents concerned and as a result, the system will retrieve many documents containing little required

information. Further, If we use more concepts representing more terms in indexing, the greater is the potentiality of false coordination in searching. Thus, in a document containing four concepts ABCD, in which A and B are related, C and D are related, there may be the possibility of retrieving documents in which A and C are related, B and D are related, and so on. Conversely, a low level of *exhaustivity* of indexing leads to *low recall* and *high precision.* Suppose, we index each input document under a central or core topic only after the adoption of the policy of indexing at the minimum level of *exhaustivity.* Obviously, the system will not retrieve the document(s) in response to a request for a topic other than the core topic and thus, will eliminate the possibility of false coordination.

On the question of *specificity*, it is to be noted that the specificity of the vocabulary is the single most important factor influencing the *precision* with which a search can be conducted. For example, we want to retrieve documents on 'POPSI' from a retrieval system. If the system includes only its broader term like 'pre-coordinate indexing', or the still broader 'indexing', we shall retrieve a large number of documents, many of which will not discuss the required topic. In such a situation, we will achieve *high recall* at the cost of *low precision.* However, we can decrease the level of *recall* by increasing the level of *specificity.* For example, if we select 'pre-coordinate indexing' we can retrieve many documents where POPSI has been discussed along with other pre-coordinate indexing techniques. But, if the system includes a more specific term 'POPSI', the number of retrieved items will come down to our specific requirements and thus, can achieve *high precision.*

An analysis of recall and precision failures is undertaken after establishing the recall and precision ratios. This *failure analysis* implies diagnosis and is intended to lead to therapeutic action. It is the most important part of the entire evaluation programme, and involves a careful examination of the documents involved, how they were indexed, the request posed to the system, the search strategy used, and the complete assessment forms of the requesters. *Failure analysis* will attribute the recall and precision failures encountered to the principal subsystems like indexing, searching, index language, and user-system interface. The most important of the possible sources of failures as identified by Lancaster in the operation of MEDLARS are illustrated below:

Table 5.2 : Sources of Failures

Factors	Recall failures	Precision failures
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 coordination Incorrect term relationships.
Indexing	Lack of specificity, Lack of exhaustivity, Omission of important concepts, Use of inappropriate terms.	Exhaustive indexing, Use of inappropriate terms.
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 combinations.
User-System	Requests more specific than actual information needs.	Requests more general than actual information needs Computer processing,
Others	Computer processing, Clerical	Clerical, Value judgment, Inevitable retrievals.

5.4.1.2 Recall and Precision Devices

Devices used to manipulate indexing files to obtain optimum recall and precision have been categorised as follows:

Table 5.3 : Devices for Optimization of Recall and Precision

Recall–oriented	Precision–oriented
Synonym control	Co-ordination
Word form control	Links
Classification, including	Roles
a) Hierarchies	Weighting
b) Lattices	Relational indexing
c) Facet analysis	
d) Semantic factoring	
e) Clumps and clusters	

5.4.2 Other Performance Measures

Apart from the recall and precision ratios, various other measures of retrieval performance have been used and suggested. We have already discussed the performance measures as suggested by Perry and Kent in the Section 5.4 of this Unit. The following measures can also be derived from the matrix representing a common frame of reference for evaluation as mentioned therein:

i) **Snobbery ratio:** It refers to the number of relevant items that are not retrieved over the total number of relevant items in the collection. It is sometimes referred to as the conditional probability of a 'miss' and is measured by c / (a+c). It is complementary of recall ratio.

ii) **Noise factor:** The proportion of retrieved items which are not relevant. It is complementary of the precision ratio and is measured by b / (a+b).

iii) **Fallout ratio:** It refers to the number of non-relevant items that are retrieved over the total number of non-relevant items in the collection. It is sometimes referred to as 'discard' or the conditional probability of a 'false drops' and is measured by b / (b+d). It measures the proportion of the non-relevant items in the file, which are retrieved by a request. The fall out thus represents the fractional recall of irrelevant items. The fallout ratio taken together with the recall ratio may provide important clues to the underlying causes of retrieval performance.

iv) **Specificity ratio:** It refers to the conditional probability of a correct rejection and is measured by d / (b+d). It is complementary to the fall out ratio.

v) **Generality number:** It expresses the number of items relevant to a particular request over the total number of items in the collection. The higher the generality number the greater the density of relevant items to total collection, and the greater this density the easier the search tends to be.

vi) **Novelty ratio:** The proportion of relevant documents retrieved in a search that are new to the requester, that is, brought to his attention for the first time by the search. This ratio is particularly appropriate in the evaluation of literature searches conducted for current awareness purposes. The *novelty ratio* can be expressed in one of two ways:

$$\frac{\text{No. of new relevant documents retrieved}}{\text{No. of relevant document retrieved}} \quad \text{OR} \quad \frac{\text{No. of new relevant documents retrieved}}{\text{No. of documents retrieved}}$$

5.4.3 Relevance

Relevance is considered as one of the important criteria underlying existing

performance measures and is most highly debated in IR research. The term *relevance* is very difficult to define. This is because there are degrees of relevance. The individual view of *relevance* led to the concept of *pertinence* or *utility*. The term *pertinence* refers to a relationship between a document and an information need, whereas the term *relevance* refers to a relationship between a document and a request statement (i.e. expressed information need). The concept of relevance needs to be viewed in the broader context of a person needing information and expressing it in the form of a request (i.e. request statement). If a document is retrieved in response to a particular request, a panel of judges may assess its relevance, but the requester can only assess its pertinence. Relevance is consensus judgment, pertinence relates to an individual judgment. Another way of looking at the matter is that a document retrieved in response to a request may be useful to the enquirer, but its utility may change. For example: If the same document is retrieved in the second search, it will have lost its utility to the enquirer. Though its relevance will not have changed, but the enquirer view of it will change. Many retrieved relevant documents may be repetitive in the second search and they do not add anything significant to our knowledge.

The degree to which a requester is able to recognise the exact nature of his information need and the degree to which his need is accurately expressed determine how successful the IR system is able to satisfy the user. In relation to most of the evaluation studies *relevance* was applied to stated requests (i.e. expressed need). But, it has now been well established that the users' requests do not reflect their information needs completely. Therefore, the current view is that the *relevance* is to be judged in relation to both expressed and unexpressed needs rather than restricting only to stated requests. But whether this would be possible is a remains the question.

The term *precision* is used very widely in the literature in preference to relevance, and the term *relevance* is used when a subjective judgment is involved. *Relevance* is used to refer to the real-life situation, *precision* is to refer to the experimental situation.

Relevance of a document to a particular request cannot be measured precisely: the relationship is subjective and equivocal rather than objective and unequivocal. Different users may make different decisions on the *degree of relevance* between a document and a request. It is also quite possible that the same user may make different decisions on *relevance* in respect of a particular request-document pair at different times. *Recall and precision ratios* measure the degree of coincidence between the user relevance assessments and the *system relevance predictions*. In a perfect search these exactly coincide. Unfortunately, such perfect searches are relatively rare. Different users have different requirements for *recall* and *precision*, and a particular individual has different requirements at different times.

The idea of measuring *relevance* spread to the Cranfield project (See Section 5.6.1 of this Unit). However, the concept of *relevance* as viewed at Cranfield was different from that of other experts. In 1967, Cuadra and Katter published a paper entitled *Opening the black box of relevance* wherein they reported that it is possible to measure relevance by obtaining higher or lower relevance scores simply by telling relevance judges how documents are to be used. For example:

a) Use in stimulating ideas, creative approaches, etc.;

b) Use in relation to a specific task; and

c) Use in preparing an exhaustive bibliography.

Lancaster opined that the *relevance* is relative and capable of being judged on some type of scale in dividing a set of documents on the basis of extent relevance [1979]:

a) clearly relevant to a particular request statement,

b) relevant to the request statement but less relevant than the first document, and

c) not relevant.

We may get disagreement among a group of judges as to which documents are relevant to a particular request statement. There does not seem to be any consensus among the judges on the question of *relevance of relevance.*

Self Check Exercises

3) What are the different criteria used in evaluating an IR system?

4) How recall and precision ratios help in measuring the performance of an indexing system?

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

.....

.....

.....

.....

.....

.....

5.5 EVALUATION METHODOLOGY

An evaluation study is conducted to determine the level of performance of the given system and also to identify those factors that are the reasons for weaknesses of the system. In other words, in evaluation an attempt is made to find out the different parameters and their interrelations with a view to assessing their contribution towards the overall performance of the system. Lancaster identifies five major steps involved in the evaluation of an information retrieval system, which are discussed below:

a) **Defining the scope of evaluation:** The first step of an evaluation study involves the designing of a precise set of questions in conformity with the evaluation objectives in order to learn capabilities and weakness of the system. The questions are usually concerned with one or more of the following:

i) Overall performance level of the system;

ii) Coverage and processing;

iii) Indexing;

iv) Index language;

v) Searching; and

vi) Input procedure and computer processing.

The purpose and scope of the whole evaluation programme are set at this step. How the evaluation study will be conducted – in a laboratory-type set-up or in a real-life situations; and the probable constraints – in terms of cost, staff, time, etc., are also considered at this stage. In fact, a detailed plan is chalked out at this stage that forms the basis for the rest of the programme.

b) **Designing the evaluation programme**: This step involves the preparation of a detailed plan of action concerning the identification of parameters and procedure for collection of data needed to answer the question(s) set in the definition of scope. The designer might need to control some of the parameters of the system while conducting an evaluation programme. The designer should point out which parameters are to be held constant during the study and how this is to be done. Data may be collected either from the existing system for an answer to the question like "What is the present response time of the system (expressed in ranges, means, median and mode)?" or by making some changes in the normal functioning of the system for an answer to the question like "What would be the effect of response time if X is applied?" Here question is answered by applying X to a representative sample of transactions and comparing the response times with the existing system.

c) **Execution of the evaluation:** It is the stage in which data (performance results like recall and precision ratios, etc.) are continuously derived in a way prescribed in the design stage. In most cases, a repeated number of observations are required to avoid sampling error and bias. Derived data are manipulated and reduced to a form suitable for interpretation and analysis so that it can answer or contribute to answering the questions set in the definition of scope. The execution of the evaluation is obviously the most time-consuming step in an evaluation study.

d) **Analysis and interpretation of results:** The success of an evaluation programme rests upon the method of interpretation of results and its accuracy. This stage should begin before the execution stage is completed. Performance results or data collected on different parameters during the execution of the evaluation programme are analysed and interpreted in this stage. Although the methodology for manipulation of the data is determined at the design stage, the evaluator might need to make some changes so as to arrive at a better conclusion. Once the data have been manipulated suitably, the evaluator gets a set of results that is to be interpreted in the light of the set of objectives. No precise guidelines for analysis and interpretation are available since they vary considerably from one evaluation application to another. Appropriate statistical techniques are applied to the analysis and interpretation of the performance results. The evaluator might need to conduct failure analysis so as to justify the results and also to suggest improvements. Lancaster mentions that the joint use of performance figures and failure analysis should answer most of the questions identified in the objectives of the evaluation.

e) **Modifying the system in the light of the evaluation results:** Finally, the retrieval system is modified, if necessary, in the light of the results of the evaluation study.

Self Check Exercise

5) What are the major steps involved in the evaluation of the performance of an IR 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.

.....

.....

.....

.....

.....

5.6 EVALUATION EXPERIMENTS

5.6.1 The Cranfield Tests

5.6.1.1 Cranfield 1

The first extensive evaluation of retrieval systems was carried out by the ASLIB with the financial assistance from the National Science Foundation at the College of Aeronautics, Cranfield, UK, under the supervision of C.W. Cleverdon in 1957. It is popularly known as the Cranfield 1 Project. Its objective was to investigate the comparative efficiency of four indexing systems: UDC, Faceted Classification, Alphabetical Subject Headings List and Uniterm Indexing system.

a) **Input Elements**

A) Four indexing systems were taken into consideration:

1) An alphabetical subject catalogue based on a subject headings list and a set of rules for construction of heading;

2) A classified catalogue based on UDC with an alphabetical chain index to the class headings constructed;

3) A catalogue based on a faceted classification (Colon Classification) and an alphabetical chain index to the class headings constructed; and

4) A uniterm coordinate index controlled by an authority list of uniterms compiled during indexing.

B) Indexers: Three indexers with varied backgrounds in proficiency in indexing were taken:

1) One indexer with experience of indexing and subject knowledge of Aeronautics;

2) One with experience of indexing but with little subject knowledge; and

3) One indexer straight from library school (fresh student) with neither indexing experience nor subject knowledge.

Each indexer was asked to prepare index entries by following the above-mentioned four indexing systems.

C) *Time-period used:* Documents were indexed 5 times using variable indexing times (2, 4, 8, 12, 16 minutes). The system was run through three phases with a view to find out whether the level of performance increases with the increasing experience of the system personnel.

D) *Materials Indexed:* 100 technical reports and articles on the general field of aeronautics and the specialized field of high-speed aerodynamics, half of which were research reports and half periodical articles.

E) *Items Indexed:* With 3 indexers X 4 indexing systems X 5 time periods X 3 runs = 18,000 items were indexed.

F) *Number of Descriptors:* The study involved the use of following types of descriptors:

1) *Alphabetical Subject Heading List:*	2864 Main headings 592 Sub-headings 1560 'See' references
2) *Universal Decimal Classification:*	2350 Class numbers 4052 Headings in alphabetical order
3) *Faceted Classification Scheme:*	1686 Class numbers

4) *Uniterm Indexing System:* 3174 Terms including
607 Proper names
267 Numerals

b) **Methodology**

1) A number of people from different organizations were asked to select documents from the collection and in each case to frame a question to which that document would be an answer.

2) The project used manufactured queries, which were formulated before the beginning of the actual search. Altogether 400 queries were formulated and all were processed by the system in each of the three phases. Thus the system worked on a total of 1200 search queries.

3) The questions were put to the indexers.

c) **Results**

1) All four systems were operating with an effectiveness that could be expressed by a recall ratio between 60% and 90% with an overall average of 80%.

2) The average recall ratios for the different systems were as follows:

- Alphabetical Index: 81.5%
- Faceted classification: 74%
- UDC Scheme: 76%
- Uniterm Indexing: 82%

Subsequently, when the facet sequence was modified the recall factor of the faceted classification scheme increased to 83%.

3) Increased time spent in indexing improved the chance of recall. Recall ratios for different timing were as follows:

Time (in minutes)	Recall (%)
2	73
4	80
8	74
12	83
16	84

The apparent drop in efficiency at the 8 minutes level was difficult to explain. Even Cleverdon himself fails to explain the same.

4) There was no significant difference in retrieving documents indexed by the three different indexers. In other words, there was no significant difference in the performance of the three different indexers.

5) Success rates in retrieving documents in broad fields of Aeronautics were noted to be 4-5% better than those on the specialized field like high speed Aerodynamics.

6) The success rate in the third group of 6000 items was 3-4% greater than the second group suggesting that the documents were indexed better in the third group. That is to say, trained indexers without indexing experience were able to perform consistently good indexing job in spite of lack of subject knowledge.

d) **Failure Analysis**

Altogether 495 failures were listed to which 526 causes could be given. The causes were categorized as:

Question failure:	17%
Indexing failure:	60%
Searching failure:	17%
System failure:	6%

e) **Significance**

The results of the Cranfield 1 test contradicted the general belief regarding the nature of information retrieval systems in many ways. Firstly, the test proved that the performance of a system does not depend on the experience and subject background of the indexer. Secondly, it showed that systems where documents are organized by a faceted classification scheme perform poorly in comparison to the alphabetical index and uniterm system. Cranfield 1 test was important in two other respects. Firstly it identified the major factors that affect the performance of retrieval systems, and secondly it developed for the first time the methodologies that could be applied successfully in evaluating information retrieval systems. Moreover, it also proved that recall and precision are the two most important parameters for determining the performance of information retrieval systems, and that these two parameters are related inversely to each other.

f) **Criticisms**

Cranfield 1 was the most significant study in the decade of 1958-68. Even then it was not free from criticisms. One of the main criticisms against this study was the artificiality without much relation to the real life situation. The questions used for the test were 'manufactured' from the source documents. It was argued by the critics that to use an article as a source document for framing a question and then to report the results of search for the same document in response to the same question is very much artificial. In a real life situation queries are not based on previous knowledge of availability of the document in the store. Another criticism was that this test identified the level of performance of the indexing systems concerned, but it did not throw any light on reasons for failure of these systems. In fact these questions were taken into consideration in the second Cranfield test.

5.6.1.2 Cranfield—WRU Test

Cranfield 1 was accompanied by other tests. An intermediate test between the Cranfield 1 and 2 conducted jointly with the Western Reserve University (WRU) needs special mention here. This was considered as a study of testing techniques rather than an evaluation of the index. The WRU system of indexing in metallurgical index was compared with a faceted system. The study took 950 documents and 114 questions based on source documents as input elements. Since this practice followed in Cranfield 1 was subjected to criticisms, this study was modified to include an exhaustive assessment of other documents for measuring recall and precision and deciding relevance. Two grades for relevance were considered–documents as relevant as the source and less relevant documents. The results of the study showed 75.8% recall and 33.7% precision for the Cranfield facet index. A critical analysis of failures placed major (67.1%) responsibility on searching and minor (18.4%) responsibility for indexing. The main causes of poor precision result of this study were grouped into two: (i) poor standard of the search programmes, and (ii) high level of exhaustive indexing. This study emphasized on the inverse relationship between recall and precision and influenced the programme of Cranfield 2 in sharpening the idea of index language devices.

5.6.1.3 Cranfield 2

The drawbacks of Cranfield 1 necessitated the conduct of further studies. The

second stage of Cranfield studies, known as Cranfield 2, began in 1963 and completed in 1966.The Cranfield 2 was a controlled experiment that attempted to investigate the components of index languages and their effect on the performance of retrieval systems. In Cranfield 2, the various index language devices were evaluated in terms of their effect on the recall and precision of a retrieval system. This study tried to assess the effect by varying each factor, while keeping the others constant. Some of the drawbacks of Cranfield 1 were eliminated in Cranfield 2 by bringing real-life situation in it and allowing feedback mechanisms between the indexers and users.

Test collection

1400 reports and research papers in the field of high-speed aerodynamics and aircraft structures were taken as input.

Query formulation

Each author of 200 selected research papers was asked to formulate questions for which he had cited the reference(s) in the paper. The authors were also asked to point out the documents that were not cited in their works but might have been relevant for the question they had formulated. The abstracts of the whole set of cited references were sent to the authors who were asked to assess the relevance of these with the questions they had formulated. The authors identified 1961 papers to be fully or partially relevant to the 279 questions obtained. Finally, 221 questions and 1400 documents were selected for the experiment. The success of the system was calculated by counting how many of the relevant papers thus assessed were retrieved by a given search.

Indexing

The documents were indexed in three different ways:

a) Each document was analysed and important concepts were selected and recorded in a natural language;

b) Concepts denoted by the single words were listed; and

c) Concepts with a weighting (ranging from 1 to 3) were combined to represent the subject contents of the documents.

Five different types of indexing languages with variations were used in this study:

- Single term index language with 8 variations like uncontrolled natural language, controlled synonyms, controlled word form, etc.;
- Simple concept index language with 15 variations by applying various controls;
- Controlled term index language with 6 similar variations;
- Title index with two variations; and
- Index language generated from abstracts with 2 variations.

Thus, 33 index languages were formed in this study by different levels of coordination of index terms.

Searching

To conduct searches 221 questions generated by the authors of the research papers were used. Relevancy of each document was ascertained for the questions by grading them from 1 to 4 in the following manner:

1) Complete answer to the question;

2) High degree of relevance;

3) Useful, providing general background of the work or dealing with a specific area; and

4) Minimum interest, providing information like historical viewpoint.

For the assessment, a single performance measure, called normalised recall, was introduced. This is a ratio of cumulated recall ratio and number of search stages involving document cut-off groups.

Results

The results of Cranfield 2 were rather unexpected because, other than corroborating inverse relationships of recall and precision, the results showed that :

1) A best performance result was obtained by the use of natural language single term index, such as Uniterm, based on words occurring in document texts;

2) With the natural language formation of groups or classes of terms beyond the stage of true synonyms or word forms resulted in fall of efficiency;

3) Use of precision devices like partitioning or intermixing was not as effective as the basic precision of coordination;

4) It was suggested that the terms taken from documents may be used successfully with minimum control in a post-coordinate index and it is helpful to eliminate synonyms. But any measures taken to control the vocabulary are likely to decrease its efficiency;

5) In the case where concepts were used for indexing, the system performance worsened with the introduction of superordinate, subordinate, and collateral classes along with the original concepts;

6) When broader and narrower terms were included along with the controlled languages of the thesaurus, the performance worsened; and

7) Index languages formed out of titles performed better than those formed out of abstracts.

Results and Conclusions

The results of the Cranfield 2 tests were unexpected because the test performing index languages were composed of uncontrolled single words occurring in documents. However, the variables used in the study were subject to criticisms. Each index language consisted of different units of words, phrases, or combinations of both. Both the documents and queries were formed in the same way. Thus the matching of questions to documents would evaluate the relative effectiveness of the languages of different specificity. Vickery comments that the measures used in the second Cranfield project do not adequately characterise those aspects of retrieval performance, those are of operational importance [1970].

Self Check Exercise

6) What were the objectives and input elements of the Cranfield 1 project?

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

.....

.....

.....

.....

.....

5.6.2 MEDLARS Test

Most of the evaluation studies were carried out on small collection. But the evaluation conducted by F.W. Lancaster [1979] on the performance of the Medical Literature Analysis and Retrieval System (MEDLARS) of the National Library of Medicine, USA during August 1966–July 1967 was based on large collection of MEDLARS database which already contained 7,50,000 records relating to medical articles on magnetic tape (8). It was first major evaluation of an operating retrieval system. From the MEDLARS tape, monthly issues of *Index Medicus* were printed and terms used to index the subject of the articles (on average 6-7 terms per article) were drawn from a thesaurus of Medical Subject Headings (*MeSH)*, which then contained about 7000 main subject headings. The study involved the derivation of performance figures and conduct of detailed of failure analyses for a sample of 3000 real searches conducted in 1966-67. The main objectives of the MEDLARS test were:

1) to study user search requirements;
2) to determine how effectively and efficiently MEDLARS was meeting the user search requirements;
3) to identify factors adversely affecting the performance of the MEDLARS; and
4) to find out the ways to improve the performance of the MEDLARS.

a) **Methodology**

At the outset, a sample work statement consisting of a list of questions to be answered in the MEDLARS study was designed. It was decided that a target of 300 evaluated queries (i.e. fully analyzable test search requests) was needed to provide an adequate test. The range of queries should, as far as possible, be representative of the normal demand covering different subjects of medical literature like diseases, drugs, public health, and so on. Representativeness was achieved by stratified sampling of the medical institutions from which demands had come during 1965, and processing queries received from the sample institutions over a 12-month period. It was also decided to include all kinds of users (academic, research, pharmaceutical, clinical, Government) for the test and they should supply a certain volume of test questions for the test. The twenty-one user groups were so selected. Some 410 queries were received from the user group and processed, and finally 302 of these were fully evaluated and used in the MEDLARS test.

Queries were submitted to the MEDLARS and on receipt of the queries, MEDLARS staff prepared a search formulation (i.e. query designation) for each query using an appropriate combination of *MeSH* terms. A computer search was then carried out in the normal way. At this stage each user was also asked to submit a list of recent articles that he judged to be relevant to his query. The result of a search was a computer printout of references. Since the total items retrieved might be high (some searches retrieved more than 500 references), 25 to 30 items were selected randomly from the list and photocopies of these were provided to the searcher fro relevance assessment. Each searcher was asked to go through the full text of the articles and then to report about each article on the following scale of relevance:

H1 – of major value (relevance)

H2 – of minor value

W1 – of no value

W2 – of value not assessable (for example, in a foreign language).

The precision ratios were calculated by using the above scale of relevance: If L items were in the sample, the overall precision ratio was 100(H1 + H2)/L and the 'major value' precision ratio were 100H1/L. It was obviously not feasible to examine the whole MEDLARS database in relation to each search in order to establish recall ratio. Therefore, an indirect method was adopted for calculation of recall ratio: Each user was asked to identify relevant items for his query before receiving the search output and then search was carried out to find out whether those items were indexed in the database and retrieved along with other items that are both relevant and irrelevant. If t such relevant items were identified by the user and available on the database for a given query, and H were retrieved in the search, the overall recall ratio and 'major value' recall ratio was estimated as 100H/t and *100H1/t1* respectively.

The next stage of the evaluation was an elaborate analysis of retrieval failures, i.e., examining, for each search, collected data concerning failures include:

a) Query statement;

b) Search formulation;

c) Index entries for a sample of 'missed' items (i.e. relevant items that are not retrieved) and 'waste' items (i.e. noise—retrieval of irrelevant items); and

d) Full text (c).

b) **Results**

The average number of references retrieved for each search was 175, with an average or overall precision ratio of 50.4%; that is, of the average 175 references retrieved, about 87 were found to be not relevant. The overall recall ratio was 57.7% as calculated by an indirect method. Taking the average search, and assuming that about 88 of the references found were relevant, with an overall recall ratio of 57.7% implies that about 150 references should have been found, but 62 were missed. However, the recall and precision ratios for each of the 302 searches were analysed and individual ratios were then averaged in the MEDLARS test. The results were:

	Overall	*Major value*
Recall ratio	57.7%	65.2%
Precision ratio	50.4%	25.7%

An elaborate analysis of retrieval failures revealed that over the 302 searches, there were 797 recall failures and 3038 precision failure. Major categories of the failures were attributed to the principal system components as shown below:

Table 5.4 : Retrieval Failures

System attribute/component	Recall failure (%)	Precision failure (%)
Index language	81 (10.2)	1094 (36)
Indexing	298 (37.4)	393 (12.4)
Searching	279 (35)	983 (32.4)
User-system interaction	199 (25)	503 (16.6)
Others	11 (1.4)	78 (2.5)

c) **Conclusions**

The results of the MEDLARS test led to a series of recommendations for the improvement of the MEDLARS performance. Some notable changes made to the MEDLARS as a result of this test include design of a new search request

form (intended to ensure that a request statement is a good reflection of the information need behind it); and expansion of the entry vocabulary and improvement of its accessibility, and the adoption of an increased level of integration among personnel involved in indexing, searching and vocabulary control devices.

Self Check Exercise

7) What was the scale used in MEDLARS test for calculating precision ratio?

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

.....

.....

.....

.....

.....

5.6.3 SMART Retrieval Experiment

The SMART retrieval system, based on the processing of abstracts in natural language forms, was launched in 1964 and Gerard Salton carried out the evaluation of various searching options offered by the SMART under laboratory condition.

The SMART retrieval system was a unique experimental environment for the development and evaluation of automated retrieval techniques. Documents were represented in the SMART system by a set of weighted terms (term vectors) and a set of documents is represented in a term assignment array. Each term was assigned a weight, which was positive, if an index term was actually assigned to a document and zero if it was not. An individual query was similarly represented as a vector of query terms. Automatic indexing using the term discrimination model carried out the creation of the document vector. In this model index terms were evaluated on the basis of their ability to increase the average dissimilarity of document description in a database. A good indexing term increased the average dissimilarity of documents in the collection; a bad one decreased it. Terms were then assigned discrimination values according to the degree to which they increase or decrease average document dissimilarity. Retrieval was based on the degree of similarity between the document vectors and the query vectors.

A number of methods were adopted for automatic content analysis of documents like Word suffix cut-off methods, Thesaurus look-up procedures, Phrase generation methods, Statistical term associations, Hierarchical term expansion, etc.

In order to develop a prototype for a fully automated information retrieval system a large number of search experiments were performed in the SMART experiment. The user environment was simulated by running iterative searches based on user feedback. The evaluation procedures of the SMART experiment was based on a pair-wise comparison of the effectiveness of two or more processing methods and a number of evaluation parameters were computed for each of the processing methods. It was followed by a comparison of the corresponding measures for two or more methods to produce a ranking of the methods in decreasing order of retrieval effectiveness. The following evaluation measures were generated by the SMART system:

i) A recall-precision graph reflecting the average precision value at ten discrete recall points — from a recall of 0.1 to a recall of 1.0 in intervals of 0.1;

ii) Two global measures, known as normalized recall and normalized precision, which together reflect the overall performance level of the system; and

iii) Two simplified global measures, known as rank recall and log precision, respectively.

a) **Methodology**

A collection of 1268 abstracts in the field of library science and documentation, mainly published in American documentation, 1963-1964, and also in some other journals was used for this experiment.

Eight different persons familiar with the subject either as a librarian or as a student of library science were asked to generate a total of 48 different search requests in the documentation field in grammatically correct and unambiguous English language. Following the receipt of the query formulation from each of the eight persons, a series of forty-eight requests were searched against a file of 1268 records, using the various search options that SMART system allows. Then, the text of the document abstracts were distributed, and each person was asked to assess the relevance of each document abstract with respect to each of his six queries. An essential element in the SMART experiment was the *relevance feedback*. If, in a preliminary output, the user can indicate which items are relevant and which are not, the system recalculate the weight of the items in the database. This is done by reducing weights associated with the characteristics of non-relevant items and increasing the weight of the characteristics associated with the relevant ones. Four sets of judgements were compared. The relevance judgement groupings were as follows:

Group	Judges	Function
A		Original group of query authors. Each person in the A group made relevance judgments for his six queries.
B		Non-author judges. Each person in the B group made relevance judgments for six queries corresponding to six different authors from the A group.
C		The document is relevant to a given query if either the A judge or B judge termed it relevant.
D		The document is relevant to a given query if both A and B judges termed it relevant.

After receiving the relevance judgements, called the 'A judgements', a second and independent set of relevance judgements called the 'B judgements', was obtained by asking each person in the test group to judge for relevance six additional queries originated by six different people. The same relevance criteria were used for the second relevance judgements as for the original ones; the only difference was that the 'A judgements' were rendered by the query authors, whereas the 'B judgements' were the non-author judgements. In order to achieve the objectivity in the experiment, the 'B judges' were not informed of the 'A judgements' obtained previously, nor was there any interaction between assessors either before or during the process of judgement.

Thus, for each of the 48 queries, a set of four different documents sets became available, each consisting of the items termed relevant by a different set of people as follows:

A set – relevance assessed by the query author;

B set – relevance assessed by outside subject expert;

C set – relevance asserted by either A or B assessor; and

D set – relevance asserted by both A and B assessor.

Three automatic language analysis procedures include in the SMART system, known as word form, word stem, and thesaurus, are furnished below:

1) Word form: Common words and final 's' endings were removed from the texts of document abstracts and queries, and weights were assigned to the remaining word forms; the resultant texts were then matched to obtain the document-query correlation coefficient.

2) Word stem: Texts were treated in the same way as word form, except that complete suffixes are removed from the text words to reduce the texts to weighted word stems; the query-document matching process remains the same.

3) Thesaurus: Each word stem produced by procedure (2) was checked with a thesaurus providing synonym recognition, and the resulting weighted concept identifiers assigned to queries and documents were compared (instead of word forms and word stems).

b) **Results**

It was found that, under normal circumstances, an evaluation of performance for a variety of processing methods required an examination of the ranking of the corresponding recall-precision curves, rather than a detailed comparison of the actual recall and precision values. From a ranking of the recall-precision graphs obtained from the several processing methods it was noted that :

a) Although the overall consistency of relevance agreements between the groups was not particularly high, the relative performance of the various retrieval methods was unaffected by changes in the relevance decisions; that is, all four sets of relevance judgements caused the same ranking of alternative search procedures. To be more specific, the word-form process was found to be less powerful than the two other procedures, and the thesaurus process was slightly better than the word stem match.

b) The best results in terms of recall and precision were obtained for the D judgements, which represented the agreements between both the A and B relevance judges; for low recall the precision was about 20% higher for D than for A, B, or C.

It was observed that the SMART evaluation output did not vary with the variations in the relevance judgments. The recall-precision output was basically invariant for the collection under study. It was concluded that if the relevance assessments by the query authors are typical of what can be expected from the general user populations, then 'the resulting average recall-precision figures appear to be stable indicators of system performance which do in fact reflect actual retrieval effectiveness'.

Self Check Exercise

8) What were the methods adopted for automatic content analysis of documents in the SMART retrieval experiment?

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

.....

.....

.....

.....

.....

5.6.4 TREC Experiment

It has been observed that almost all the earlier evaluation experiments were based on a small data set and unrelated to the real-life problem. The major problem for IR researchers was to base the evaluation experiments on large test collection to match a real-life situation with an infrastructural support for conducting test on them. Under such circumstances, TREC (Text REtrieval Conference) series of experiments in information retrieval, funded by DARPA (Defence Advanced Research Project Agency, Department of Defence, USA) and operated by the NIST (National Institute for Science and Technology, USA), was launched in 1991 in order to enable IR researchers to scale up from small collection of data to larger experiments. The TREC series of experiments has drawn attention of the LIS professionals all over the world since its inception and has shown that significant research results can be obtained through international efforts and collaboration.

a) Objectives

The objectives of the TREC experiments are to :

1) encourage IR evaluation experiments on large test collections;
2) enable the exchange of ideas among industry, academia, and government in an open forum created for the purpose;
3) facilitate the speedy transfer of technology from research labs to commercial products by demonstrating substantial improvements in retrieval methodologies on real-life problems; and
4) develop new evaluation techniques applicable to current retrieval systems and to make them available for use by industry and academia.

b) Scope

A wide range of information retrieval strategies was tested in different TREC experiments (i.e. from TREC 1 in 1992 to TREC 12 in 2003). Some notable examples are:

a) Boolean search;
b) Statistical and probabilistic indexing and term weighting strategies;
c) Passage or paragraph retrieval;
d) Combining the results of more than one search;
e) Retrieval based on prior relevance assessments;
f) Natural language-based and statistically based phrase indexing;
g) Query expansion and query reduction;
h) String and concept-based searching;
i) Dictionary-based searching;
j) Question-answering;
k) Content-based multimedia retrieval.

c) Structure of TREC Experiment

The two-fold information retrieval activities—i.e. (1) *Core* (i.e. Main activity), and (2) *Tracks* (i.e. Subsidiary activities) have been recognised by the TREC for the purpose of experiment. The following figure shows the structure of the TREC experiments:

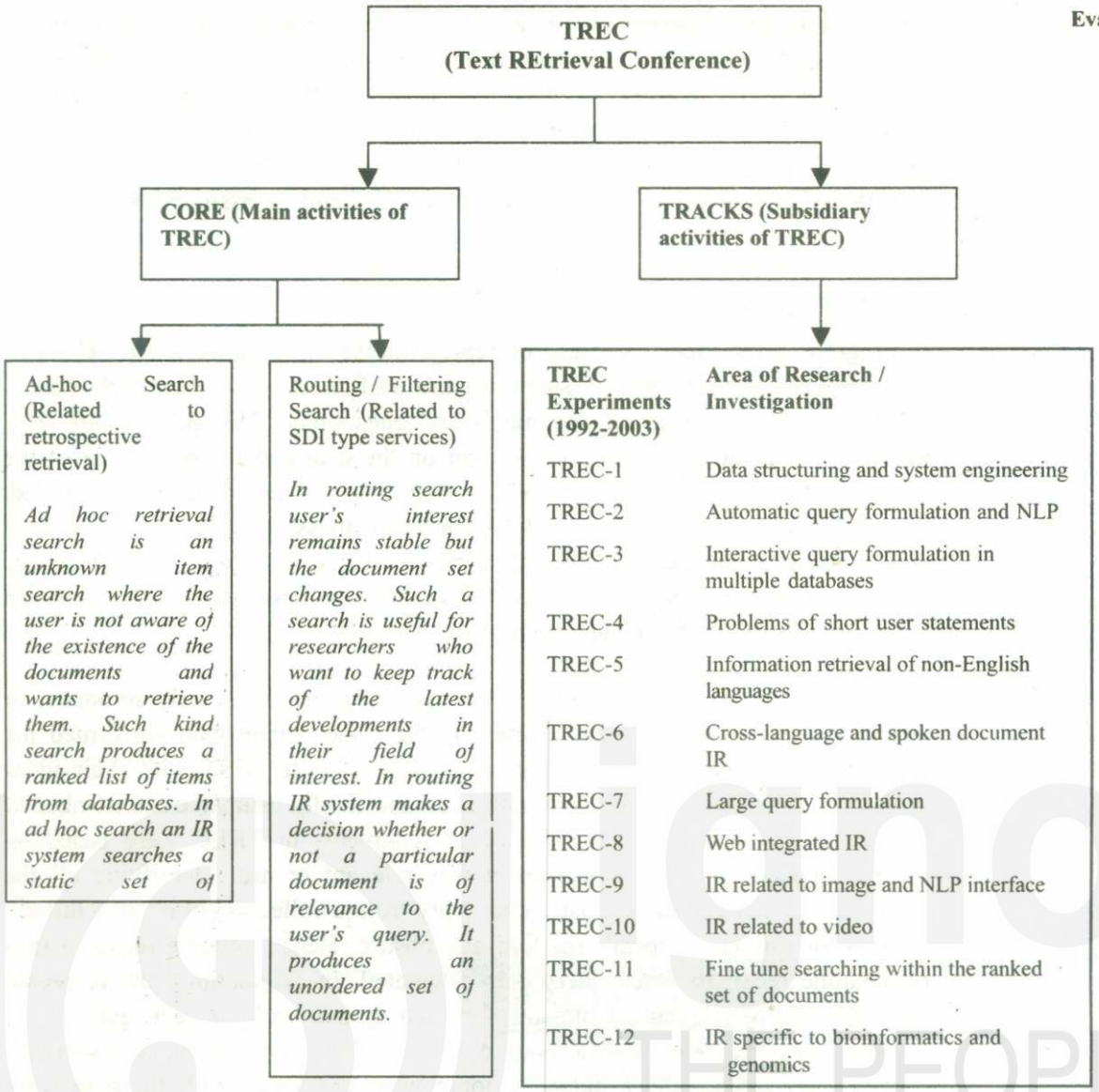


Fig. 5.2 : Structure of TREC experiments

Core or main activity has been divided into (i) *Ad-hoc search* (i.e. retrospective retrieval), and (ii) *Routing or Filtering search* (i.e. SDI). The ad-hoc task investigates the performance of the system that search a static set of document—from a database or from the Web—in response to a question and produces a ranked list of items. In TREC jargon, an information need is termed as a *topic,* and the data structure submitted to a retrieval system is called a *query. Routing or filtering search* is needed either (a) by the researchers to keep track of the latest development in their field of interest, or (b) by the analysts to monitor news feed for items of interest on a particular subject. This type of search produces an unordered set of documents as opposed to a ranked list. Here, the retrieval system makes a decision whether or not a particular document is of relevance to the given user's query.

A *track* acts as an incubator for a new area of experiment. The first running of a *track* usually defines the problem and creates the necessary infrastructure, including test collections, evaluation methodology, etc. to support experiment on the specific task. TREC experiments have also become evolutionary in the sense that new tracks have been added as and when new problems areas are encountered by the researchers.

d) TREC Collection

The document collection of TREC (known as TIPSTER collection) reflect diversity of subject matter, word choice, literary styles, formats, and so on. The primary

TREC collection consists of about 2 gigabytes of data with over 8,00,000 documents. The document collection used in various tracks depends on the needs of the track and availability of data. The primary TREC document sets consist mostly of newspaper or newswire articles. Some government publications such as Federal Register, patent documents and computer science abstracts are also included. Each document is encoded in SGML and is assigned a document number.

e) Methodology

A program committee consisting of representatives from government, industry and academia was formed to supervise the TREC activities. For each TREC, NIST provided a test set of documents and questions. TREC participants were allowed to run their own retrieval system on the data and returned a list of the retrieved top ranked documents to NIST where the individual results were pooled, the retrieved documents were judged for correctness, and finally, the results were evaluated. The TREC cycle ends with a workshop— a forum for participants to share their experiences. The TREC workshop is organised every year and latest one was the 12th in the series, held at NIST in 2003.

Topic statements (i.e. query statements), which may be either fully automatic or purely manual or mixed, were created by the same person who performed the relevance assessment for that topic. Assessors came to NIST with their ideas for topics, searched the document collection and finally estimated the number of relevant documents per topic. Relevance judgments in TREC were based on binary decision—whether a document was relevant or not. Identifying all the relevant documents against a particular query from a collection of about a million records for estimating recall was logically absurd. Hence, relative recall instead of absolute recall for each query was calculated. In calculating relative recall, 100 top-ranking relevant documents for each system and for each query were combined together to produce a 'pool' of relevant documents. Then, each system's retrieved documents for a query or topic were compared with these relevant documents. The proportion of the 'pool' of relevant documents that a single system retrieved for a query or topic is relative recall. Output lists of 100 top-ranking documents from each participating research team were sent to NIST where the 100 top-ranking documents for each topic from all the participating research teams were merged into a single set and then handed over to the assessor for relevance evaluation. Most of the relevant documents were found by pooling all the results from all the participating teams. The results for each system were then analysed to generate various performance measures.

The ad-hoc retrieval tasks in TREC were evaluated using package called tree-eval, which reported about 85 different numbers for a run, including recall and precision measures at various cut-off points and a single value summary measure for recall and precision. The recall—precision curve (a plot of precision as a function of recall) and mean average precision (mean of the precision obtained after each relevant document is retrieved) were the most commonly used measure to describe TREC results.

f) Results

The TREC series of experiments has produced very significant and interesting results. The findings of each experiment (along with specific reports) appear regularly on TREC Website (http://trec.nist.gov). Some important findings of the TREC experiments are furnished below:

i) Use of 'pooling' to produce the sample results for relevance judgments was

found to be more than adequate for test collection;

ii) Automatic construction of queries from natural language queries seems to work well, or better than, manual construction of queries, which is encouraging for groups supporting the use of simple natural language interfaces for retrieval systems;

iii) Significant improvement of retrieval performance over TREC 1 was observed in TREC 2 in respect of the routing task with the increase of documents per topic (200 per topic to 1000) and database size (1 gigabyte to 3 gigabytes);

iv) Level of performance was same in spite of differences experimental designs. For examples: some groups generated queries automatically from the topic statements while others generated the queries manually; relevance feedback was not incorporated in many systems; the computer platform used ranged from PCs to a supercomputer;

v) Differences in precision-recall curve was minimal; and

vi) Although the precision-recall results were similar, there was a large scatter in the actual documents retrieved.

g) **Criticism**

The main criticism against TREC centres round on the methodological issues. TREC experiments were carried out within the traditional laboratory paradigm, which is very difficult to relate to users browsing information on the Web as opposed to traditional library setting.

Self Check Exercise

9) Mention the different information retrieval strategies that were tested in the TREC series of experiments.

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

.....

.....

.....

.....

.....

5.7 SUMMARY

This Unit presents a holistic view of the evaluation of IR systems. The indexing system is a major component of IR system. In the context of information processing and organisation, evaluation of indexing system is thus most crucial for improving the efficiency of the system. This Unit thus lays major emphasis on the evaluation of indexing system. It starts with the discussion on different levels of evaluation like system effectiveness, cost effectiveness and cost-benefit evaluation, and different evaluation criteria consisting recall, precision, fallout, specificity, novelty, relevance, etc. Major steps involved in the evaluation of the performance of an information retrieval system have been discussed. Important evaluation experiments—such as, Cranfield 1, Cranfield-WRU test, Cranfield 2, MEDLARS test, SMART Project and TREC have also been discussed in terms of their objectives, scope, methodology, results, significance and criticisms.

5.8 ANSWERS TO SELF CHECK EXERCISES

1) By *effectiveness* we mean the level up to which the given IR system attains its stated objectives, whereas *efficiency* refers to how economically the given IR system is achieving its stated objectives. In order to know the *effectiveness* of the system we measure the extent to which the given IR system can retrieve relevant documents while withholding non-relevant documents. But, we measure only the cost factors like response time, user effort, cost involved per search, etc. for determination of the efficiency of the system.

2) An information retrieval system can be evaluated at the following levels:

   a) System effectiveness;

   b) Cost effectiveness; and

   c) Cost-benefit evaluation

3) Different criteria used in evaluating an IR system are :

   a) Recall;

   b) Precision;

   c) Response time;

   d) User effort;

   e) Form of presentation; and

   f) Collection coverage.

4) The term *Recall* relates to the ability of the system to retrieve relevant documents. Recall ratio is nothing but the proportion of relevant documents retrieved. This ratio helps in measuring the extent to which the retrieval of relevant documents occurs in a given system. The formula for the calculation of *Recall ratio* is:

$$\frac{\text{Relevant retrieval}}{\text{Total relevant}} \times 100$$

On the other hand, *precision ratio* is the ratio between the relevant retrieved and total retrieved documents. It measures how precisely an indexing system functions. The formula for the calculation of the *precision ratio* is:

$$\frac{\text{Relevant retrieval}}{\text{Total retrieval}} \times 100$$

5) The following major steps are involved in the evaluation of the performance of an IR system:

   a) Defining the scope of evaluation;

   b) Designing the evaluation programme;

   c) Execution of the evaluation programme;

   d) Analysis and interpretation of results; and

   e) Modifying the system in the light of the evaluation results.

6) The objective of the Cranfield 1 project was to investigate the comparative efficiency of four indexing systems: UDC, Faceted classification, Alphabetical subject heading list and Uniterm indexing system. Input elements for this project consist of :

   a) Four indexing systems as stated above;

   b) Three indexers with varied backgrounds;

c) Variable indexing time periods;

d) 100 micro documents on the general field of aeronautics and the specialized field of high-speed aerodynamics;

e) 18,000 indexed items (100 micro documents × 3 indexers × 4 indexing systems × 5 time periods × 3 runs); and

f) Descriptors.

7) In the MEDLARS test, the user was asked to characterize each retrieved item in a randomly selected sample output of 25 to 30 items by using the following scale: H1 – of major value; H2 – of minor value; W1 – of no value; W2 – of value unknown (because in foreign language). From this assessment, precision ratios were calculated: If L items were in the sample output, the overall precision ratio was 100(H1 + H2)/L; the 'major value' precision ratio was 100H1/L.

8) A number of methods was adopted for automatic content analysis of documents in the SMART retrieval experiment. Some notable examples include:

a) Word suffix cut-off methods,

b) Thesaurus look-up procedures,

c) Phrase generation methods,

d) Statistical term associations,

e) Hierarchical term expansion and so on.

9) A wide range of information retrieval strategies was tested in different TREC experiments—that is, from TREC 1 in 1992 to TREC 12 in 2003. Some notable examples are:

a) Boolean search;

b) Statistical and probabilistic indexing and term weighting strategies;

c) Passage or paragraph retrieval;

d) Combining the results of more than one search;

e) Retrieval based on prior relevance assessments;

f) Natural language-based and statistically based phrase indexing;

g) Query expansion and query reduction;

h) String and concept-based searching;

i) Dictionary-based searching;

j) Question-answering; and

k) Content-based multimedia retrieval.

5.9 KEYWORDS

Clumps : A group of words that tend to 'adhere'. Various criteria can be used to define the boundaries of a clump. In general, each word is associated with the others as a group above some given threshold. The formation of word classes is likely to be based on more than simple concurrence. Instead, two words are considered to be related if they co-occur more frequently than one expects them to co-occur probabilistically.

Clustering : A group formation mechanism that combines a number of related or associated terms. Four kinds of group may be found in clustering: strings, stars,

cliques and clumps. String occurs when term A is strongly associated with term B, term B with term C, and so on. In practice, strings tend to form loops fairly quickly: term A->B->C->D->E->A. Stars are found when one term is equally strongly related each to other. Cliques occur when a set of terms are strongly related each to the other. Clumps are weaker form of clique, in which a term is related to one or more of the others in the clump, but not necessarily to all.

Cost effectiveness : It is the evaluation in terms of how to satisfy user requirements in the most efficient and economical way. A cost effectiveness evaluation relates measures of effectiveness to measure of cost. It takes into considerations unit cost per relevant citation retrieval, per new relevant citation retrieval and per relevant document retrieval.

Cost-benefit : It is the evaluation of the worth of the system by relating the cost of providing the service to the benefits of having the service available.

Elimination factor : The proportion of non-retrieved items (both relevant and non-relevant) over the total items in the collection.

Evaluation : It refers to the act of measuring the performance of the system in terms of its retrieval efficiency (ease of approach, speed and accuracy) to the users, and its internal operating efficiency, cost effectiveness and cost benefit to the managers of the system in order to ascertain the level of its performance or its value.

Exhaustivity : It is the measure of the extent to which all the distinct topics discussed in the documents are considered.

Failure analysis : A diagnostic procedure, which involves the analysis of recall and precision failures encountered to the principal subsystems of an IR system like indexing, searching, index language, and user-system interface. It consists of a careful examination of the documents involved, how they were indexed, the request posed to the system, the search strategy used, and the complete assessment forms of the requesters.

Fallout ratio : It is the proportion of the irrelevant items in the file, which are retrieved by a request. The fall out thus represents the fractional recall of irrelevant items. It has been referred to as the conditional probability of 'false drops'. It is also called 'discard'.

Generality number : It expresses the number of items relevant to a particular request over the total number of items in the collection. The higher the generality number the greater the density of relevant items to total

collection, and the greater this density the easier the search tends to be.

Noise factor : The proportion of retrieved items those are not relevant. This factor is considered as the complement of the precision ratio.

Novelty ratio : The proportion of relevant documents retrieved in a search that are new to the requester, that is, brought to his attention for the first time by the search.

Omission factor : The proportion of non-relevant items retrieved over the total number of non-retrieved items in the collection.

Performance Curve : A performance curve is nothing but a line representing graphically a series of variable performance points in terms of recall and precision ratios affected by search strategy at varying levels, from very broad to very specific. This curve expresses the inverse relationship between recall and precision.

Resolution factor : The proportion of total items retrieved over a total number of items in the collection.

Semantic factoring : Involves the analysis of a word into largest number of ideas implied. During the 1950s a team at Case Western Reserve University worked on a system of analysis known as semantic factoring. The objective was to break down every concept into a set of fundamental concepts called semantic factors. Because of their fundamental nature, there would only be a limited number of these factors. For example, word such as 'urinalysis' factors into 'urine' and 'analysis', whereas 'magneto-hydrodynamics' factors into 'fluid flow', 'magnetism' and 'electrical conductivity'.

Specificity : It is measure of the degree of preciseness of the subject to express the thought content of the document.

System effectiveness : It is the evaluation of the system performance in terms of degree to which it meets the users' requirements. It considers the users' satisfaction and is measured by determining the utility to the users of the documents retrieved in answer to a user query. It takes into consideration cost, time and quality criteria.

Usability : The value of references retrieved in terms of such factors as their reliability, comprehensibility, currency, etc.

User effort : It refers to the intellectual as well as physical effort required from the user in obtaining answers to the search requests. The effort is measured by the amount of time user spends in conducting the search or negotiating his enquiry with the system.

5.10 REFERENCES AND FURTHER READING

Chakraborty, A.R. and Chakraborty, B. (1984). *Indexing: principles, processes and products*. Calcutta: World Press.

Chowdhury, G.G. (2004). *Introduction to modern information retrieval*. 2nd Ed. London: Facet Publishing.

Cleverdon, C.W. (1962). *Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems*. Cranfield: College of Aeronautics.

Foskett, A.C. (1996). *Subject approach to information*. 5th Ed. London: The Library Association.

Ghosh, S.B. and Satpathi, J.N. (eds.) (1998). *Subject indexing systems: concepts, methods and techniques*. Calcutta: IASLIC.

Gopinath, M.A. (1999). *Evaluation of information storage and retrieval (ISAR) systems*. In: MLIS-03, Block 3, Unit 12 course materials. New Delhi: Indira Gandhi National Open University.

Jones, Karen Sparck. (ed.) (1981). *Information retrieval experiment*. London: Butterworth.

Lancaster, F W. (1979). *Information retrieval systems: characteristics, testing, and evaluation*. 2nd Ed. New York: John Wiley.

Salton, G. and McGill, M.I. (1983). *Introduction to information retrieval*. New York: McGraw-Hill.

Sarkhel, J.K. (2001). *Information analysis in theory and practice*. Kolkata: Classique Books.

TREC. Available at http://trec.nist.gov/pubs.html.

Vickery, B.C. (1970). *Techniques of information retrieval*. London: Butterworth.

ignou
THE PEOPLE'S
UNIVERSITY