INFORMATION PROCESSING AND RETRIEVAL

Dr. B. R. AMBEDKAR OPEN UNIVERSITY
HYDERABAD
2004

COURSE TEAM

Sl.No.	*Course Editor/Writer*	*Course Units*
1.	Dr. V. Chandrasekhar Rao Dr. B.R.A.O.U., Hyderabad.	*Editor*
2.	Sri T. Damodaram Dir. Oil Seeds, Research, Hyderabad.	1-3
3.	Sri M. Ramachander P.S. Telugu University, Hyderabad.	4 & 6
4.	Prof. A.A.N. Raju (Retd.) Osmania University, Hyderabad	5
5.	Dr. Binayak Patnaik Lingaraj Law College, Berhampur (Orissa)	7
6.	Dr. V. Chandrasekhar Rao Dr. B.R.A.O.U., Hyderabad.	8, 15,16
7.	Sri P. Ratnakumar (Formerly) ICRISAT, Hyderabad.	9-12
8.	Sri S. Subba Rao CLRI, Chennai (Tamil Nadu)	13
9.	Sri P. Divakar CCMB, Hyderabad.	14

Cover Design : CHANDRA

Dr. B.R. Ambedkar Open University
Hyderabad.

First Published 1999, Reprint 2001 2004

This text forms part of an Open University Course.

Further information on Open University programmes may be obtained from the Director, (Academic Dr. B.R. Ambedkar Open University, Prof. G. Ram Reddy Marg, Road No. 46, Jubilee Hills, Hyderabad - 500 033. (A.P.) INDIA.

Lr.No.1026/Dr.BRAOU/DMP/PTG/F.NO15/JO.13/Copies :100, 2004- 05/Dated 11-11-2004
Printed at M/s. Shri Gopala Kwik Graphics, Hyderabad – 39.

COURSE - 03 : INFORMATION PROCESSING AND RETRIEVAL

The Course Information Processing and Retrieval conforms to the syllabus of the Master of Library and Information Science (MLISc) offered by Dr B R Ambedkar Open University. As you are familiar with the structure of the programmes, syllabus and the courses, the print material developed by the Open University has some distinct features. For the sake of convenience, the syllabus is divided into block, each of which comprises a number of units. Each unit generally covers a specific area of the subject. The units are prepared by specialists in accordance with the format so designed to enable you read and understand them without much difficulty. Each unit begins with the structure (contents list) and statement of its aims and objectives, followed by an introduction to the content of the unit. The subject contents of a unit are divided into sub-themes and are numbered up to three levels for easy reference and comprehension. Each unit end up with Let Us Sum Up, Glossary, Assignments, References and Further Reading, and Model Examination Questions.

Ever since the information and communication technologies have entered the library and information fields, there has been a complete metamorphosis in their structure, functions and services. Computer stored and produced documents, indexes, bibliographies, databases, and information services have become the order of the day. The whole work of information storage and retrieval can be broadly divided into three areas: 1) Content analysis of documents, 2) Representation of the content in a suitable form of records and creation of file, and 3) Actual retrieval of information or document surrogates. The first area is completely an intellectual work. The content is represented through classification numbers or index terms as document surrogates. The details of the documents are stored in databases as records in a standardised format for easy retrieval and exchange of information among different systems. Access to enormous quantities of information available in different databases (online or CD-ROMs) poses certain challenges to the users. Various search strategies have been used to access and retrieve the information from the databases, including the Internet based ones. The on-going research in information retrieval leads to designing Library Expert Systems; where the intelligent computer systems have built-in expertise of human specialists is available in problem solving situations. Expert system intermediaries and expert systems-based search engines are becoming the potential aids for database searching. All these developments have been introduced briefly in this Couse.

The Course aims to provide an overview of the various methods used in information storage and retrieval in libraries and information centres. The specific objectives of the Course are -

- to provide an overview of the bibliographic description and to describe the standards used in bibliographic record formats of the print and non-print materials;
- to acquaint the learners with various general and special classification schemes (with emphasis to Universal Decimal Classification), subject indexing systems (with emphasis on PRECIS and POPSI) and information retrieval thesauri;
- to familiarise the learners with the information storage and retrieval systems, their objectives, methodology, and experiments and case studies conducted for their evaluation.
- to provide an understanding on information access from online and CD-ROM databases, search strategies used for information retrieval, and the use of expert systems.

CONTENTS

BLOCK - I : BIBLIOGRAPHIC DESCRIPTION OF PRINT AND NON-PRINT MATERIALS

Bibliographic Description is a process of identifying and recording certain characteristics of a document for its identification, content specification, location, access and dissemination purposes. The bibliographic elements (i.e., author, title, edition, publication details, physical description, series, standard numbers, etc.) are chosen to provide information adequately about the document that is not in hand. Some of the bibliographic details also serve as document surrogates or access points in library catalogues. There have been several attempts to codify the rules for entry elements (headings) and document description to ensure uniformity, consistency and efficiency in cataloguing right from Anthony Panizzi of British Museum Library to the present times. An overview of the bibliographic description has been provided in the first unit of this block.

The need for standardization of bibliographic elements in descriptive cataloguing was felt after the entry of computer technology and facilities for exchange of information among different systems. Machine readable bibliographic record formats, such as US-MARC, UK-MARC, UNIMARC, UNISIST Reference Manual, CCF and Indian Standard IS:11370-1985, have been evolved. These standard formats have been discussed in Unit-2.

Bibliographic materials, in formats other than print, have been gaining importance as effective storage, information retrieval and dissemination media in recent decades. There are multitudes of formats, in which non-print materials are available - microforms, sound recordings, visual media, videograms, magnetic and optical discs, digital video discs, etc Using AACR-2R, cataloguing of some of the major formats of non-print materials have been described and worked out examples have been provided in Unit-3.

In this block, you find these three units under the following titles:

Unit-1: Bibliographic Description - An Overview

Unit-2: Standards for Bibliographic Description

Unit-3: Bibliographic Description for Non-Print Materials

UNIT - 1 : BIBLIOGRAPHIC DESCRIPTION - AN OVERVIEW

Structure

1.0 AIMS AND OBJECTIVES

This unit aims to give an overview of the bibliographic description and the efforts made through out the world to achieve standardisatioon towards in bibliographic description.

After reading this unit, you should be able to

- describe the concept of bibliographic description
- appreciate the need for standards for bibliographic description
- analyse the efforts for developing standards for bibliographic description
- acquire the knowledge of international standards for cataloguing/bibliographic description
- identify the rules for bibliographical references and citations.

1.1 INTRODUCTION

Library collection includes well known traditional forms of printed material (books, monographs, serials, maps etc.) and non-print media (microforms, microfilm, microfiche, film strips, slides, sound recordings, magnetic disks, CD-ROMs etc.). There has been a steady growth of non-print materials in the past two decades. The non-print materials keep appearing in modern libraries and becoming more and more useful as sources of information. The Collection of a library should be listed in an orgnised and purposeful way to recall the items when required. This externalization of recall procedure is done through a library catalogue.

The library catalogue records certain selected and agreed features of a document in a pre-determined pattern. While cataloguing, a document is identified first in a bibliographic description and then is made accessible in different approaches by adding a number of explicit access points such as heading for author, title and subject. The catalogue records built up in a systematic pattern to serve as a permanent representative of documents which may not always be at hand in the library. The catalogue is, thus, a permanent list of a library which contains records that are prepared according to a set of rules or code of rules, the code being a standard for representing a document by a number of significant data elements. The standard or code is intended to fulfil two objectives of a catalogue record i.e., identification by bibliographic description and providing access points or headings. Review of catalogue codes reveal that the catalogue code makers have not given equal importance to these two important objectives. Some cataloguers have laid emphasis on headings to build a particular form of a library catalogue while others envisaged standard bibliographic description. It was a precursor for efficient cataloguing for producing a variety of catalogues/ or indexes.

1.2 CONCEPT OF BIBLIOGRAPHIC DESCRIPTION

Bibliographic description is a process of identifying and recording certain characteristics of a document for its identitification, content specification, location, access and dissemination purposes. The bibliographic elements are chosen to provide information adequately about the

doucment that is not in hand. The bibliographic description is therefore, a process of recording the bibliographic details about the document so that to serve as a surrogate i.e., a finding list. In other words bibliographic description is a document surrogate. Bibliographic description is generally used in document linked operation for instance bibliographic reference or a citation of a text. Bibliographic description also helps in overcoming problems in acquisition, organisation and use of documents in a library. Though the subject information, content compatibility are found in both print and non-print materials, the physical description of these materials changes. Authoritative description of these documents is very important.

1.3 DEVEPMENTS IN RULES FOR BIBLIOGRAPHIC DESCRIPTION

Bibliographic description calls for certain standardization that will ensure uniformity, consistency and efficiency in the process. The main areas generally considered for standardization are : 1) headings, 2) bibliographic description itself, and 3) areas or attributes of a bibliographic record. Standards of form and presentation of bibliographic record are essential pre-requisites to achieve international compatibility of bibliographic record. Some of the important early attempts to prescribe certain norms/rules to achieve uniformity and consistency in cataloguing will be presented to you.

The era of scientific cataloguing based on code of rules for cataloguing to guide the inventory and finding functions of a library catalogue began with the publication of famous the "Ninety One Rules" developed for the British Museum by Anthony Panizzi and his associates. Though only fourteen rules were given for description of the books, the "Ninety one Rules" had a strong influence on the subsequent catalogue rules in Europe and America. Charles Coffin Jewette codified 39 rules in 1850 in a different orientation than that of "Ninety one Rules". The principal emphasis of Jewette code was on the principles of uniformity and consistency in cataloguing. His concept of compilation of a general catalogue from separate uniform bibliographic units was very similar to International Standard Bibliographic Descriptions of the 1970s, which you will be studying in detail in due course. Jewette had given cataloguing rules a logical structure: first bibliographic description next access points through headings and cross references. The first definite code to mark the era of systematic cataloguing appeared in 1876 with the publication of Charles Ammi Cutter's *'Rules for printed Dictionary Catalogue"* which in later editions was titled as *"Rules for Dictionary Catalogue"*(RDC).

The RDC by including basic definitions of cataloguing terms, choice of headings in entries and details of descriptive cataloguing "appear to have been the chief source of latter codes in English language". Attempts to draw such codes were also made in other countries for example the *"Rules for alphabetical catalogue"* (RAK) in Germany. It is very comprehensive code of rules, covering filing and transliteration rules, lists of abbreviations as well as rules for headings and bibliographic descriptions. Further, the RAK represents a major revolution in cataloguing in that, for the first time, the concept of corporate authorship was introduced and the mathematical rather than grammatical title was accepted. Thus several attempts were made between 1841 and 1900 to develop codes for cataloguing. During this period Library Association of the United Kingdom and Library of Congress of USA also brought out their codes.

However, in the beginning of the next century further developments were made to evolve a code widely acceptable through out the globe. Professional organisations showed keen

interest in these endeavors. On the suggestion of Melvil Dewey, American Library Association and Library Association (Great Britain) made united efforts to produce *Anglo-American Code* in 1908 with a view to establish uniformity in cataloguing. ALA on its own revised 1908 code in 1941 and published *"ALA Catalog Rules: Author and Title"*. It contained two parts. Part 1 Entry and heading and Part 2 Description of book. ALA further undertook revision of part 1 of 1941 code and revision of Part 2 was deferred. The Library of Congress brought out *"Rules for Descriptive Cataloguing in the Library of Congress"* in 1949. The same was proposed to be the substitute for Part 2 of the 1941 code.

Meanwhile, in India Dr S. R. Ranganathan has made significant contributions to cataloguing. He designed code for classified catalogue and dictionary catalogue. His contributions to cataloguing are unique in the sense that he built theoretical foundations for practical cataloguing on the basis of canons, besides five laws of Library Science and general laws such as Law of Symmetry, Law of interpretation , Law of impartiality, Law of Parsimony etc. These general laws, norms and canons provide basic guidance to develop, design and oraganise aspects of bibliographic description.

Despite many productive efforts in cataloguing, an international agreement on cataloguing principles underlying cataloguing practice posed certain problems from time to time. This necessitated IFLA sponsored *International Conference on cataloguing principles* in 1961 in order to arrive at certain basic agreements in cataloguing. The generally agreed twelve statements of principles, popularly referred to as "Paris Principles" limited itself to the standardization of headings and the organization of alphabetical catalogue. The Paris Conference also recommended for a study on the impact of electronic machinery and of mechanical procedures on cataloguing rules and paved the way for future compatibility in standardization of bibliographic description in automated environments.

1.4 STANDARDS FOR BIBLIOGRAPHIC DESCRIPTION

The developments that took place in Library and Information Science from the early 1960s revolutionised the concept, scope and purpose of bibliographic description. The exponential growth of literature resulted in the development of different tools such as bibliographies, union catalogues, indexing and abstracting services to ensure effective bibliographic control. This gave the impetus to develop adequate formats with standard bibliographic description. The centralised and cooperative cataloguing efforts further intensified the need to have a standard bibliographic format that will promote the resource sharing among libraries. The efficient information retrieval system must have a format of bibliographic description that suits its requirements. This has also accelerated the move to modify and redesign the existing bibliographic formats compatible to the new information environment. The micrographic, computers and telecommunication technologies brought a sea change in the library environment. The changes in the pattern of library and information services with the introduction of new technologies rendered the traditional manual oriented bibliographic description standards and practices inadequate and thus the need for newer bibliographic standards.

It is with this background the *International Meeting of Cataloguing Experts* (IMCE) was convened by IFLA in Copenhagen in 1969. The attention of cataloguers shifted to the creation of a standard pattern for the bibliographic description. The experts agreed to work for an international standard framework for bibliographic description with all the required descriptive data that would serve the needs of both cataloguers and bibliographic agencies. A Working Group consisting of A. J. Wells and Michael Gorman was set up for this purpose.

1.4.1 International Standard Bibliographic Description (ISBD)

The Working Group set up to draw detailed provisions on the Standard Bibliographic Description (SBD) submitted the draft international standard bibliographic description to IFLA Committee on Cataloguing in 1971. The preliminary version of the draft International Standard Bibliographic Description (ISBD) was published in 1971 and with revision it was first published as a standard edition of ISBD(M) in 1974. ISBD became an important programme of IFLA since then. The main purpose of ISBD Programme is to provide a uniform descriptive framework for all types of library materials that will serve varied bibliographic uses. Comprehensiveness fixed order of data elements and use of specific punctuation as delimiters between bibliographic elements are the distinctive characteristics of ISBD format. Keeping in view the peculiarities of different forms of materials such as serials, cartographic materials, non-book materials, printed music, antiquarian, audio visual materials, computer files etc. the ISBD(M) followed by publication of series of specialized ISBDs, viz., ISBD (S), ISBD (CM), ISBD(NBM), ISBD (PM), ISBD (A), ISBD (AVM) and ISBD (CF). An integrated general format for all types of documents, called General International Standard Bibliographic Description (ISBD(G) was also brought out to bring harmony among various ISBDs and to remove incompatibilities among them.

The contribution of ISBDs to the standardization of bibliographic record is regarded as a significant achievement for the following reasons:

1) ISBD facilitates records from different sources interchangeable that is to say exchange of data through networks.

2) ISBD framework assists interpretation of records across language barriers since bibliographic items in each record can be easily identified through specialized punctuation and its place in the record.

3) ISBD also assists in the conversion of bibliographic records to machine readable form.

In ISBD, traditional data elements of bibliographic description are separated and grouped into eight distinct areas, each of them is marked by specific punctuation symbols and set out in a sequence which is easy to be analysed and understood by the user and the machine alike. ISBD helps standardization of cataloguing rules and manual format for machine readable bibliographic records.

1.4.2 Standards for Bibliographic Record Formats

The Standard record formats for bibliographic description are required for international compatibility and promotion of procedures and practices for machine readable databases with a strategy for development towards universal bibliographic control. These exchange formats in connection with bibliographical control facilitates comprehension and transfer of bibliographic data between bibliographic agencies at national and international levels across the countries. The growth of ISBDs and the similar formats in various countries renewed interest in a common interchange or communication formats. The exchange of bibliographic data in different situations for example networking among a group of libraries, use of bibliographic utility services, downloading of data from on-line databases, cooperative bilateral exchange

agreements etc., have resulted in the concept of a communication format. The purpose of a communication format is to promote free-flow of information by facilitating the exchange of bibliographic records created in a database to a machine readable form. Library of Congress was the first to design and experiment on such a format viz., Machine Readable Catalogue Record (MARC) format for the purpose of communicating bibliographic information to a large number of libraries. Improvements led to the development of MARC II. With growing enthusiasm many national bibliographic agencies developed their own individual MARC formats based on US-MARC format. The need for the establishment of international format for the exchange of bibliographic data, IFLA took the initiative to develop international MARC, resulting the development of UNIMARC for monographs and serials in 1977 and the second edition was brought out in 1989. UNESCO developed Common Communication Format (CCF) as a universal exchange format for bibliographic records.

1.5 RULES FOR ENTIRE BIBLIOGRAPHIC RECORD

1.5.1 Content of bibliographic record

A bibliographic record includes sufficient details about the item being described for its identification. Such details are called "data elements". In other words a data element is a word or group of characters representing the distinct unit of bibliographic information and forming part of the bibliographic description of an item. Elements providing physical description of an item for example pagination, size etc., other elements such as author, title, subject, place of publication, name of publisher, ISBN etc. are considered data elements. In a bibliographic record, description of data elements is governed by rules provided in a 'Catalogue code' or 'Cataloguing Rules'. The rules in the Catalogue Code/Rules cover all aspects of the description of a bibliographic item in a record. Although there is no universally accepted set of rules for bibliographic description we find a number of national and international codes of cataloguing and rules for special applications. Most of the Codes/Rules are generally in conformity with the general cataloguing principles or 'Paris principles' that were developed under the auspices of IFLA at the International conference held in 1961 in Paris. This is particularly true with the AACR. The Paris principles have had a profound influence on the later developments in catalogue code revisions.

1.5.2 Anglo-American Cataloguing Rules (AACR)

The AACR is widely used in English-speaking countries and more or less adopted translations are in use in many other countries. In 1967 AACR was published in two versions: North American and British texts. Since there were no international guidelines for the development of rules for description, AACR used LC Rules for Bibliographic Description as the basis for developing rules for description of monographs, serials and non-book materials. The logical treatment of rules in AACR and its emphasis on the conditions of authorship rather than types of work were considered to be a great improvement over the previous codes. Keeping in view the need for international cooperation and standardization in cataloguing, the AACR was revised and the second edition was published in 1978 as a single text. AACR (2nd ed) is divided into two parts: Part 1 : Description and Part 2 : Headings, Uniform titles and References. Part 1 is based on ISBD (G) for special types of materials and Part 2 is based on the 'Paris Principles'. A revision was brought out for the AACR second edition in 1988.

1.5.3 Rules for Description of Non-Book Materials

The inadequate treatment of rules for non-book materials in general cataloguing codes necessitated the development of special rules applicable for non-book materials. Concerted efforts were made to provide standard rules for bibliographic description for NBMs since the 1950s. Supplementary Rules for "LC Rules for Descriptive Cataloguing in Library of Congress" were published for Phonocards in 1952; Motion pictures and Filmstrips in 1953; Pictures, designs and other two dimensional representations in 1959. Canadian Manual for Non-Book Materials was published by Canadian Library Association in 1970 and updated in 1979. The British Library Association Media Cataloguing Rules Committee published "Non-Book materials cataloguing rules: Integrated code for practice and draft revision of British Text Part III" in 1973. This is Popularly known as LANCET rules designed to be used in conjunction with AACR. IFLA's standard bibliographic description ISBD (NBM) appeared in 1977. Detailed rules for description of NPMs are given in AACR (Second edition 1988 Revision).

1.5.4 Specialized Rules for the Entire Bibliographic Record

The AACR, RAC and some other general cataloguing rules have been conceived in the first instance for organisation of materials in libraries. Rules have also drawn up primarily intended for broad group of information services particularly abstracting and indexing services. In some cases the rules specifically suitable for a more narrowly defined category of bibliographic material such as government, scientific and technical reports. These manuals of rules are not specifically concerned with creation of headings they intend to combine rules for bibliographic description with a record structure in which it allows automatic generation of a variety of headings and other required access points by automatic machine processing. We shall be discussing some of the specifically developed manual of rules for bibliographic description.

1.5.5 UNISIST Reference Manual

MARC format was considered possible solution for machine processing of bibliographic records in libraries. The automation of secondary information services i.e. Abstracting & Indexing Services necessitated standardized record format. A consequent effort of UNESCO, the first edition of Reference Manual for Machine Readable Bibliographic Description (UNISIST RM) was published in 1974. Subsequently the second edition was brought out in 1981. Organizations were making use of the Reference Manual as a source for bibliographic description as well as guidelines for a machine readable exchange format.

1.5.6 Guidelines for Processing of Documentary Literature

The "Guidelines for Processing of Documentary Literature" is a manual for bibliographic description developed specifically for Abstracting and Indexing services in Federal Republic of Germany. It is useful for both manual computerized systems. It covers the description of all kinds of bibliographic materials though main emphasis is on the printed materials. It contains rules for choice of headings, form of citations and filing. The rules have drawn up based on ISBD(M), ISBD(S), COSATI, Chemical Abstract service and INIS rules for bibliographic description Guidelines for Descriptive Cataloguing of Reports. These rules for descriptive cataloguing of scientific and technical rules were first published in 1966 by the US Committee on Scientific and Technical Information (COSATI) of the Federal Council of Science and

Technology. The COSATI rules were updated and published in 1978. These rules were intended to libraries and information centres who process/exchange bibliographic information of scientific/technical reports especially on magnetic a tape or on disc. The rules though cover minimum data elements are very specific to identification retrieval of the material. A number of premier organisations in USA viz. Defense Documentation Centre, NATIS and NASA are using it for bibliographic description of scientific and technical reports.

1.5.7 Rules for Bibliographic References

Bibliographic references which appear in bibliographies, lists publications, references, foot notes in a book etc. are generally prepared by authors/editors/compilers and contain bearest minimum bibliographic data elements necessary to identify and locate the items cited. Many countries have established the national standards for bibliographical references to ensure uniformity in description, in their respective countries. The International Standard Organization has published "ISO 690-1975: Documentation - Bibliographical References - Essential and Supplementary Elements" Similar to ISO 690, American National Standard Institute brought out 'ANSI 239.29-1977. In India, The Indian Standards Bureaux (formerly Indian Standards Institute) has published IS:2381 in 1963 and revised it later.

1.6 RULES FOR PARTS AND PARTICULAR ATTRIBUTES OF BIBLIOGRAPHIC RECORD

Rules for the parts and particular attributes of bibliographic record specifically include headings and entry elements, name headings, uniform headings for titles, etc. The following sub-sections deal with the efforts towards the standardisation of these elements.

1.6.1 Standards for Headings and Entry Element

In cataloguing this is the most difficult area to arrive at standardization, as there are strong divergent national and local traditions in respect of choice rendering. Nevertheless, some efforts towards standardization have taken place mostly in English speaking countries. Though the concept of main entry heading played major role in cataloguing, with the advent of computerization of cataloguing and indexing the concept of equal value access point gained more emphasis and found useful approach than that of main entry heading or entry. However, in manual systems and single entry listings such as bibliographies and citations the choice of main entry heading remains important.

1.6.2 Name Headings

The established national form of personal name is generally accepted internationally. IFLA has issued guidelines on national usage for entry in catalogue of names of persons and supplementary volume covering more countries is under preparation. Divergent practices also exist with regard to the choice and rendering of corporate author as heading. Some organisations for example INIS have established their own authority lists of corporate headings for use in their information services. IFLA has made some efforts to cover the special problems represented by governments and their agencies as authors. IFLA prepared a list of uniform headings for higher legislative and ministerial bodies in European countries. Similar effort to publish for African countries is also initiated.

1.6.3 Uniform headings/Title headings

When a work does not have recognizable personal or corporate authors, or the work has many authors and authorship is diffuse, the cataloguing codes suggest that the work should be entered under a uniform heading or under title. You will find typical example of such works in religious and liturgical works, serials etc. The IFLA has prepared a list of uniform title headings for anonymous classics in medieval European literatures. The classic example of title entry headings is serial or periodical. Though identification and description of serial is well known by title, some titles are not so specific for example transactions, bulletin etc. and further, you find changes often occur in the name of the title of serial. The International Serials Data System (ISDS) provides a unique identification of all serials by means of a standard form of 'key title' associated with its International Standard Serial Number (ISSN) based on the guidelines of UNESCO manual of guidelines for ISDS.

1.7 Standard Numbers

International standard numbers have been developed for unique numbering of monographs and serials. Ten digit International Standard Book Number (ISBN) is assigned for each monograph. As said earlier eight digit International Standard Serial Number (ISSN) is assigned for serial title. CODEN is also alternatively used for ISSN as unique identifying number internationally and particularly in USA. CODEN consists of characters and numerals. The standard "ISD3388-1977: Patent Documents: Bibliographic References: Essential and Complementary Elements" is intended for the construction of unique international identification numbers for patents. International Organisation for Standardization (ISO) initiated to develop an international numbering scheme for technical reports based on United States "ANSI 239.23-1974: Standard Technical Report Number (STRN)". US National Technical Information Service (NTIS) also use unique report number, which is widely used through out the world. Two draft international standard numbering system viz. "ISO/DIS 3901: International Standard Recording Code"(ISCR-1975) and "ISO/DP 5956:International Standard Record Number" were developed for sound recordings.

1.8 LET US SUM UP

Let us now try to summarize what we have discussed in this unit. In spite of several efforts for the last one and half a century since the publication of the famous "Ninety one Rules" we could not find unified and internationally accepted cataloguing rules for the bibliographic record. Variations in different national requirements impede the possibility of developing a unified code/rules. However, the profound unifying influence of Paris "Statement of principles" should never be underestimated especially in the area of choice and form of headings, though the creation of headings or the concept of main entry remains subject of debate. A number of national and international cataloguing codes/rules including AACR were developed under the influence of 'Paris Statement of Principles".

The area where significant progress has been made is that of standard bibliographic description especially with the introduction of IFLA's International Standard Bibliographic Descriptions. The rapid and generally enthusiastic acceptance of the ISBDs in many national and other libraries is amazing. Peculiarities of different forms of documents necessitated

development of specialized ISBDs viz., serials, antiquarian, non-book materials, audio-visual materials, computer files, printed music etc. ISBD format, due to its elaborate display format with a complex system of punctuation, could not impress Abstracting & Indexing services, which relies on type of a bibliographic description which allows identification of discrete bibliographic data elements and their sub-components to make possible generation of a variety of headings and display formats through computer processing. To meet the requirements of the Abstracting & Indexing services and other information services Unesco developed the Reference manual. The need for international compatibility in exchange of information in machine readable formats and networking of computers necessitated development of standard bibliographic record formats viz., US-MARC, UK-MARC, UNIMARC and CCF formats.

The standard numbering system viz. ISBN and ISSN, standard codes for names of countries and languages, abbreviations of typical words in bibliographic references and standard abbreviated titles of serials are widely used in bibliographic descriptions.

1.9 GLOSSARY

Access point : A name, term, code etc., under which a bibliographic record may be searched and identified.

Bibliographic elements: Elements in a document representing author, title, publisher, series, ISBN etc.

Dictionary catalogue : A catalogue in which all the word entries are arranged in alphabetically in one sequence.

Paris principles : The twelve statement of principles so named and were drawn up at the International Conference on Cataloguing Principles, 1961, Paris. In cataloguing Author/title entry should be based on these principles. The subject matter of the twelve principles are: 1) Author and title entries, 2) Functions of the catalogue, 3) Structure of the catalogue, 4) Kinds of entry, 5) Use of multiple entries, 6) Function of different kinds of entry, 7) Choice of uniform heading, 8) Single personal author, 9) Entry under corporate bodies, 10) Multiple authorship, 11) works entered under title and 12) entry word for personal names.

1.10 REFERENCES AND FURTHER READING

BOWERS, Fredson. *Principles of bibliographic description.* Winchester, St. Paul Bibliographies, 1986.

VERONA, Eva. *A decade of IFLA's work on the standardization of bibliographic description.* (1980)

VICKERY, B. C. "Bibliographic description, arrangement and retrieval". IN: *Introduction to information science/* Saracevic Tefko. New York, 1970. pp. 428-35.

1.11 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) Define bibliographic description in your own words

2) List out the different ISBD's developed for different types of materials

3) What are the rules and guidelines which are concerned with bibliographic description of non-book materials?

4) What are the unique identification numbers, which are use for monographs and books?

II SHORT NOTES

a) ISBN

b) Paris Principles

c) UNISIST Reference Manual

UNIT-2 : STANDARDS FOR BIBLIOGRAPHIC DESCRIPTION

Structure

2.0 AIMS AND OBJECTIVES

The aim of this unit is to introduce to you the standards for bibliographic description of various documents.

After reading this unit, you will be able to:

- explain the meaning of bibliopgraphic record format
- trace the developments for standardised bibliographic record formats
- describe the development of ISBD format
- discuss various machine readable cataloguing formats, viz., US MARC, UK MARC, UNIMARC;
- follow Common Communication Format and UNISIST Reference Manual.

2.1 INTRODUCTION

Library catalogue is a bibliographic record representing documents by a number of significant data elements. Two important processes in cataloguing that is 1.Identification by bibliographic description and 2.Providing access oints for author, title, subject content etc., have not had equal consideration in discussions of catalogue code makers. The early cataloguers viz., Anthony Panizzi and C. A. Cutter gave priority headings and organizational aspects of a catalogue while relegating establishement of standard bibliographic description to second place. Some others for instance French Government in 1791, Jewett in 1850, Crestadoro in 1856, the Prussion Instructions in 1899, the Machine Readable Cataloguing (MARC) project and other forms of computerised cataloguing, International Standard Bibliographic Descriptions (ISBDs) and lastly AACR rules emphasized on preparation of uniform or standardised bibliographic records.

The standardisation of bibliographic description is aimed at compatability and promotion of procedures and practices with a strategy for development towards universal bibliographic control which is a precurser to universal access to publications. Application of computers in the development of bibliographic databases has raised the hope of developing universal bibliographic system through the participation of several national and international organisations.

We shall examine the various efforts in designing and developing standard bibliographic record formats.

2.2 BIBLIOGRAPHIC RECORD FORMAT

A bibliographic record describes a document. It comprises four major components.

(1) A description of physical entity or document itself in terms of such bibliographic elements or data elements as author, title, publisher etc.

(2) Selection of elements from the description of the document to use as approach points through which the record can be retrieved.

(3) A unique record identifier of the document.

(4) Descriptors representing the subject content of the document drawn from the authority lists such as subject heading lists or thesaurus or natural language terms drawn from the document itself.

The Bibliographic description is itself built on several data elements. A data element is a distintive unit of information forming part of the bibliogaphic description and having a specific functional relationship with the content of the document, e.g., author, title, publisher, edition etc. The place where data element is to be recorded is known as field. Each field in a printed record such as a catalogue or bibliography or a documentation list is fixed in a format to identify data elements separately within a record so that each field is recognised and distinguished among the different fields using various clues i.e., position, indention, print font or type face, punctuation etc.

A machine readable bibliographic format serves a similar purpose. It is like a cabinet with compartments designed to fit the data in such a way that each data element is identified and can be manipulated, recalled, compared, stored, printed etc., by a computer program. Any machine readable record format refers to the structure, content and coding of the record. Structure provides the frame work for incorporating fixed length as well as variable length fields or combination of both types in a record. Content refers to the data contained in a record in different fields. Coding is the digital representation of the characters.

2.3 STANDARDS FOR BIBLIOGRAPHIC RECORD FORMAT

2.3.1 Need and Purpose

The Purpose of standard record format in connection with bibliographical control facilitates comprehension and transfer of bibliographic data between bibliographic agencies at national and international levels. It is very essential to design and develop a standard record format uniformly acceptable to all agencies involved in exchange of information.

2.3.2 Background

Before we examine the developments in machine readable record formats it is required to take note of the standardization attempts in the context of manually prepared bibliographic records.

The first standard set of "Ninety one Rules" was drawn up in 1841 by committee consisting of Anthony Panizzi for British Museum catalogue. This set of rules has had strong influence on subsequent rules for cataloguing and bibliographic description in English speaking countries. Only 14 rules (18-31) out of ninety one rules specifically address bibliographic description of books. Subsequently, Jewett's far reaching plan in 1852 to construct a general catalogue step by step from separate uniform bibliographic units had the genus of co-operative cataloguing and to the concept similar to the ISBD's in the 1970s.

United efforts were made to produce Anglo-American Code in 1908 and 1949 with a view to establish uniformity in cataloguing. With the publication of American revision of 1949 code attention was given to the principles underlying the cataloguing practice. This idea resulting IFLA sponsored International Conference on Cataloguing Priciples (ICCP) in 1961.

The generally agreed statement of twelve principles limited itself to the standardization of headings and the organization of alphabetical catalogue which has been in the foreground for over 100 years. The Paris Conference in its resolutions also prepared the way for the future by recommending a study on the impact of electronic machinery and of mechanical procedures on cataloguing rules.

The International Meeting of Cataloguing Experts (IMCE) convened by IFLA in Copenhagen in 1969 marks the end of an era in which discussions mostly centred on questions concerning choice and form of catalogue headings and the attention of cataloguers shifted to the creation of a standard pattern for the bibliographic description. The experts in the meeting agreed to work for creation of a frame work for bibliographic description that would serve the needs of both cataloguers and bibliographic agencies. In an important policy statement the IMCE recommended the creation of a system for the international exchange of information in which the national agencies would prepare standard bibliographical descriptions of their own publications and distribute them through the medium of cards or machine readable records.

2.4 INTERNATIONAL STANDARD BIBLIOGRAPHIC DESCRIPTION (ISBD)

As a follow up to IMCE resolutions a draft International Standard Bibliographic Description (ISBD) was published in 1971, which was subsequently endorsed by ISO. ISBD become an important programme of IFLA. The main purpose of ISBD Programme is to provide a uniform descriptive framework for all types of library materials that will serve varied bibliographic uses.

The distinctive characteristics of ISBD format are:

1) Its comprehensiveness

2) Its fixed order of data elements

3) Its use of punctuation as delimiters or dividers between the different bibliographic elements.

ISBD has distinctive punctuation that separates the varied data within the bibliographic record. The spaces left before and after the special punctuation (except comma and point) aid in identification of each element.

The first ISBD standard developed in 1974 and revised in 1978 intended for the description of monographs (ISBDs). It was followed by a series of specialized ISBDs for various forms of documents as given below.

Serials ISBD (S) - First edition 1977.

Cartographic materials ISBD (CM) - First edition 1977.

Non-Book materials ISBD (NBM) - First edition 1977.

Printed music ISBD (PM) - First Published 1980
Second revision 1991.

Antiquarian ISBD (A) - First Published, 1980;
Revised 1990.
Second revision 1991.

Computer parts (Formerly known as analytics): ISBD(CP):1992

Computer files ISBD (CF) : 1990

An integrated general format for all sorts of documents, called General International Standard Bibliographic Description (ISBD(G) was also brought out.

The eight areas of data elements and the order of fields as recommended in ISBD(G) format are shown in Appendix-1

The contribution of ISBDs to the standardization of bibliographic record is regarded as the greatest achievement for the following reasons:

(1) ISBD facilitates records from different sources interchangeable that is to say exchange of data through networks.

(2) ISBD framework assists interpretation of records across language barriers since bibliographic items in each record can be easily identified through specialised punctuation and its place in the record.

(3) ISBD also assists in the conversion of bibliographic records to machine readable form.

ISBD helps standardization of cataloguing rules and manual format for machine readable bibliographic records. ISBD Provided a 'step closer to realization of IFLA's programme of Universal Bibliographic Control (UBC). In fact, IFLA's International Office for UBC administer the ISBD programme.

2.5 COMMUNICATION OR EXCHANGE FORMATS

The developments of ISBDs and the growth of similar formats in various countries renewed interest in a common interchange or communication format. The exchange of bibliographic data in different situations for example networking among group of libraries, use of bibliographic utility services, downloading of data from online databases, cooperative bilateral exchange agreements etc., have resulted in the concept of a communication format. A communication format is a sequential structure involving a single file independent of both the file structure used and the computer software. Its aim is to make the record structure as clearly visible as possible so that a receiving information system is able to easily convert the incoming data into its own internal format. The purpose of a communication format is to promote the free flow of information by facilitating the exchange of bibliographic records created in one database to a machine readable form.

Communication or exchange formats consists of three components.

1) The carrier or physical record structure comprising rules for the arrangement of data to be exchanged on a computer medium including specifications for the physical medium which may be paper tape, magnetic tape, disk, or even transitory medium as in the case of online exchange.

2) Codes are content designators to identify different data elements in the record. Content designators comprise field tags, indicators and sub-field codes.

3) Rules for the formulation of different data elements. This component is very closely tied up with the second component above. The data elements separately identified by the codes in the exchange format have to be identified not only in terms of content but also in form if the records are to be suitable for use by another agency.

Effective exchange of bibliographic data between agencies can be accomplished only if the records of the agencies conform to all the three components as said above.

In case of the first component, there exist an accepted International Standard (ISO 2709). Since internationally different cataloguing rules prevail, agreement on second and third components have not been arrived at so far. While general carrier structure is accepted as standard (ISO 2709) there are many differences in implementation.

2.6 MACHINE READABLE RECORD FORMAT

Library of Congress was instrumental in designing and experimenting on Machine Readable Catalogue (MARC) record format.Council on Library Resources granted aid to this project. This experiment was aimed at communicating bibliographic information to a large number of libraries. Before the commencement of MARC Pilot Project in 1966, there were no MARC formats available.The four major objectives of MARC pilot project are:

(1) to set a standard for communication of machine readable data between libraries.

(2) to structure the data in a format that would allow modification and adoptation within a given library.

(3) to include data elements needed by most libraries.

(4) to design a format usable on a variety of computers and related equipment.

The Pilot Project format was first developed for use by participating libraries. This format subsequently called MARC I which restricted itself to book material. In the light of experiences gained from the evaluation of MARC I, the need was felt for a standard communication format for interchange of bibliographic data across national and international organisations.MARC II was designed as one format structure capable of containing bibliographic information for all forms of material such as books, serials, maps, music, journal articles etc.

2.6.1 US MARC

There are three main components that make the present US MARC record structure:1.Leader 2.Directory 3.Variable fields.

Leader	Directory	Variable fields

Figure-1: US MARC structure

Leader provides general information about the record. Directory is the guide to the contents in a record. Variable fields contain data of particular record and they are of two types one Variable Control fields and the other Variable data fields.

1) Leader: The leader is a fixed length of twenty four characters (positions 0-23). It provides specific information for processing a particular record data such as total length, status, type of material, base address of data and encoding levels (full, minimal, complete or incomplete).In short, Leader allows a program to recognise a particular record and establish how it is to be processed.

2) Directory: The Directory is an index to the locations of variable control fields and variable data fields within a record. The Directory enables the programme to locate specific fields efficiently which required in the processing of information.It is automatically created by the MARC record processing computer programme.It contains one entry for each field in a record showing the field label(tag), its length and its starting position relative to the first field in the record.The Directory ends with a field terminator.

Tag	Length of Data field position	Starting character position	Field terminator

Figure-2: The Directory structure

3).Variable fields: Directory is followed by variable control fields and variable data fields.The variable control fields are numbered 001 to 009 and variable data fields numbered 01X to 09X for various fields.Two kinds of content designations are used within variable data fields.They are: 1.indicators and 2. sub-field codes.A tag is assigned to each variable data field and the tag is stored in the directory.For example, 100 is the tag for personal name main entry , 110 is the tag for corporate author main entry and 250 is for edition. Indicators precede the field and supplies additional information. Sub-field codes are special characters like $ or T and called as delimiters.

2.6.2 UK MARC

BNB showed active interest in the possibilities offered by LC MARC pilot project. This led to MARC in UK in 1966. BNB MARC project was launched in 1967 with financial assistance by Office of Scientific and Technical Information (OSTI). Preliminary version of BNB MARC II was issued in 1968 as MARC Record Service Proposals. Later BNB MARC II specifications was published in 1969. British MARC format was developed for the production of printed BNB and other library related purposes. The advent of BLAISE

(British Automated Information Service) and the publication of AACR II and the need to use BNB MARC format as a standard for recording different materials from other sources as well as distribution of centralised records made necessary changes in the BNB MARC format and ultimately leading to the publication of UK MARC format. The structure of the UK MARC format is given below:

| Record label | Directory | Control fields | Variable data fields |

There is some divergence between US MARC and UK MARC formats in respect of terminology and other aspects.Adoption of different texts of AACR caused some difrferences between the two formats. The second edition of AACR had its impact on the second edition of UK MARC manual which appeared in 1980 primarily in terms of coordinated treatment for materials of all kinds.Contrarily, US MARC format never confined to AACR only but reflected various cataloguing codes applied in American cataloguing practices. As said earlier there were merely differences of terminlogy between two formats such as "leader" in US MARC versus "Record label" in UK MARC format or "Bibliographic level" in US MARC versus "Class of record" in British format.Besides, there were also differences between these formats in respect of calculation of fixed field positions , fixed length data elements and control numbers used. For example, the record control number used in US MARC is the LC card number whereas in UK MARC it is ISBN.

The need for standardisation was felt as there were differences in these formats as well as 30 and odd US MARC based formats developed throughout the globe. The tripartite format structure, i.e., carrier format, content designators and data elements was proposed for acceptance as an American standard by Z39 Committee as ANSI Z.39.2-1971 standard for the format for interchange of machine readable bibliographic information on magnetic tapes. This standard was subsequently accepted with modifications by ISO as ISO 2709-1973:Documentation Format for Bibliographic Information Interchange on Magnetic Tapes.Second revision of this standard was published in 1981.The ISO 2709 describes a generalised structure or a frame work specially designed for communication between data processing systems and not intended for use as a processing format within systems. Although this standard was designed specifically for recording and processing data on magnetic tapes, its structure can be used for other data carriers (Appendix-2).

ISO 2709 structure specifies that a bibligraphic record should comprise four segments and the generalised structure of the records is shown below schematically.

| Record label | Directory | Variable data fields | Record separator |

The record label(also known as leader) and directory are both control segments which are used to process data contained in the third segment i.e the variable field segment. A record is of variable length.It ends with record terminator or separator.

2.6.3 UNIMARC

National MARC projects were rapidly increased in number world -wide after the advent of USMARC and UKMARC formats and to name a few CANMARC in Canada; INTERMARC in France; IBERMARC in Spain; INDOMARC in Indonesia; AUSMARC in

Australia; THAIMARC in Thailand; SINGMARC in Singapur; JAPMARC in Japan. Despite development of several national MARC formats the only area of standardisation in these formats was record structure which was in conformity with ISO 2709 structure. There were wide variations among different formats. The content designators used by national MARC systems were not standardised. The tags, indicators, sub-field codes, occurrence identifiers and data element identifiers varied very much among national formats. Consequently, tailor-made computer programmes had to be written by each national implementing agency to access MARC data of every other national agency, thus defeating the very purpose of standardisation of record formats.

This situation led IFLA to assume the responsibility for establishing an international standard for content designators. IFLA Working Group on Content Designators was formed. The Group identified that divergence among national formats was due to lack of standards in cataloguing practices, differences in subject systems and headings in name authority files, language differences etc. and these differences work against compatibility of data exchanged between national liraries. The Working Group proposed in 1973 a SUPERMARC based on the ISBD areas.This later became MARC International Format (MIF) from which UNIMARC was developed. UNIMARC was designed to serve as an exchange format which would necessitate to develop only two conversion programmes i.e., one from national format to UNIMARC and the other from UNIMARC to national format.UNIMARC format was first published in 1977 and then again in 1980.UNIMARC handbook was later published in 1983 for the guidance of users in different implementing agencies. All these documents were combined into a UNIMARC Manual in 1987.UNIMARC is maintained by IFLA UBCIM office.

Structure of UNIMARC

The four important elements underlying the UNIMARC structure are :

1) Modularity in the form of different blocks
2) Support for ISBD
3) Support for different forms of documents collected in libraries
4) Support for cataloguing of different forms of material at all levels

In UNIMARC, the 3 digit tags fall into the following special block structure.

	Data types	Block number	Field tags
1.	Identifiers (Bilographic item and record)	0	0xx
2.	Coded information block including language and country.	1	1xx
3.	Descriptive information	2	2xx
4.	Notes	3	3xx
5.	Linking entry block (Links to other blocks)	4	4xx
6.	Related entry block (Variant titles)	5	5xx
7.	Subject analysis block	6	6xx

8.	Intellectual responsibility	7	7xx
9.	International use	8	8xx
10.	National use block	9	9xx

"x" substitutes for number.For example 000-009,001-199 etc. ISBD is specifically accommodated in Block 2. The descriptive details could be standardised despite differences in cataloguing practices. UNIMARC currently provides content designators for records of variety of textual and non-textual materials.

2.6.4 UNISIST Reference Manual

While MARC format was being evolved for machine processing of biblographic records in libraries, secondary information services such as abstracting and indexing publications were also being automated. This trend necessitaed a standardised record format.

UNESCO together with ICSU-AB, The Abstacting Board of International Council of Scientific Unions set up a Working Group including members from ISO, FID, IFLA, INIS and OECD. A first draft of communication format was developed in 1971 and tested at the University of Sheffield. In the light of these experiments, a few changes were made to the Manual. The data elements definitions were enhanced to ensure standardised bibliographic descriptions and with this the manual began resembling cataloguing rules. The tagging system was changed to an alphanumeric system to avoid possible confusion with the existing numeric system in MARC formats. The first edition of Reference Manual for Machine Readable Bibliographic Description was published in 1974. Organizations were making use of the Reference Manual as a source of bibliographic description as well as guidelines for a machine readable exchange format.

In the Reference Manual, A tags are used for bibliographic description and B tags are used for subject indication. Users of UNISIST Reference Manual could use Z tags for local information. Two character sub-field codes are used in the Manual, the first being @ and the second may be a number or letter. Each field uses 2 indicators.

Besides, the differences in tagging system, significant differences are found between MARC and Reference Manual. Indicators not used in MARC are usually set at blanks rather than as zero as in UNISIST Reference Manual. The users of UNIMARC were libraries and national bibliographic agencies while that of Reference Manual were indexing and abstracting services.

2.6.5 Common Communication Format (CCF)

Following the advent of ISBDs, many international bibliographic record formats were developed including ISO's wellknown format for bibliographic information interchange on magnetic tape (ISO 2709). Due to the proliferation of bibliographic record formats around the world on one hand and lack of compatability among them, led to convening of the International Symposium on Bibliographic Exchange Formats by UNESCO within the framework of General Information Programme(PGI). This Symposium was organised by the UNISIST International Centre for Bibliographic Descriptions (UNIBID) in cooperation with ICSU-AB, IFLA, and the ISO in Taormina, Sicily in April 1978. The purpose of the Symposium was to deliberate the desirability and feasibility of establishing maximum compatability between existing

bibliographic exchange formats. As a result of the Symposium an ad-hoc Group on the Establishment of Common Communication Format was formed by UNESCO/PGI.

The Ad-hoc Group made the following decisions before developing Common Communication Format (CCF).

A. Structure of the new format would conform to ISO 2709.

B. CCF core record would comprise limited number of mandatory data elements identified in a standard manner which are essential to bibliographic description.

C Core record would be augmented by additional optional data elements, identified in a standard manner.

D. Standard technique would be devised for accommodating levels, relationships and links among bibliographic entries.

Besides, CCF would act as a bridge between major international exchange formats.

Initially, the Group studied the data elements of the six standard formats viz., UNISIST Reference Manual, UNIMARC, ISDS Manual, MEKOF-2, ASIDIC/EUSIDIC/ICSU AB/NFAIS Interchange specifications and the USSR - US Common Communication Format. Commonly used data elements in these six formats were identified. These common elements formed the core of CCF.

The Ad-hoc Group also developed a technique to show relationships between two bibliographic records and between elements within the same bibliographic record. This apart, the concept of record format and method for designating relationships between records, segments and fields were also developed.

The first edition of CCF was published in 1984 and the second edition in 1988 with certain modifications based on experiences of bibliographic agencies from several countries. The publication entitled "Implementation Notes for users of Common Communuication Format (CCF) was brought out as the guide for the benefit of CCF users. UNESCO/PGI brought out 3rd edition of CCF in two volumes. They are volume 1 CCF(B) : Bibliographic and volume 2 CCF(F): Factual.

CCF is useful to exchange bibliographic records among various libraries as well as abstracting and indexing services. It permits a bibliographic agency to use a single set of computer programe to manipulate bibliographic records received from different information systems. It also provides a list of useful data elements which forms as the basis of a format for an agency's own bibliographic data base. However, CCF is not meant for internal storage and processing of data in an institution, since processing formats vary from institution to institution and also within the same institution.

CCF is specifically designed for retrieval and output within an institution. We find CCF is different from other formats in two ways. Firstly, it neither includes its own cataloguing rules nor recommends any particular cataloguing code or set of rules oriented towards any specific type of output format. Secondly, it identifies and defines relartionship between the data elements and their respective content designators independent of category of items or

different types of material. In other words it outlines a specific system of content designators for bibliographic records pertaining to all forms of documents or bibliographic entities.

Record structure of CCF

The record structure of the CCF is a specific implementation of the ISO 2709. Each CCF record comprises four parts.

a) Record lable
b) Directory
c) Dat field
d) Record separator

The record lable comprises 24 characters to provide parameters to process the record and the directory 14 characters in five parts. Data fields consists of indicators, subfield identifier, subfields and field separator. The final character in a CCF record is the record separator. You will find the detailed schematic structure of the CCF record in Appendix-4. A CCF record may contain bibliographic descriptions of more than one item. Description of each bibliographic record occupies a single record segment. The primary item occupies the primary segment and the others the secondary segments. These segments are linked through two types of relationships, viz., vertical relationship (a monograph and a chapter in it) and horizontal relationship (Change of titles of a periodical)

Data elements in CCF II

In CCF, data elements have been listed in numerical order of the three digit tags. The list of data elements is given in Appendix-5. In order to import/export of the data from different data bases uniform tag numbers as assigned in CCF II have to followed.

2.6.6 Indian Standard IS: 11370-1985

Standardisation of bibliographical record format has not received due attention in India. Bureau of Indian Standards (Formerly Indian Standards Institution) made some efforts in this direction and published Indian national MARC format IS:11370-1985 titled "Guide for data elements and record format for computer based bibliographic description for different kinds of documents" in July 1986. Structure of this format is in conformity with ISO 2709-1981 format and in addition provides an illustrative list of content designators as a means of identifying data elements pertaining to any bibliographic entitity.

Structure of IS: 11370

The general structure of the bibliographic record format is given below. A more detailed structure is shown schematically in Appendix-5.

LEADER

DIRECTORY

DATA FIELDS

RECORD SEPARATOR

2.7 LET US SUM UP

The need for standardisation in bibliographic records was expressed as far back as 1841 when a standard set of "Ninety one Rules" was drawn up for the British Museum catalogue. Subsequently, Jewett in 1852 'sown the seeds'for the development of cooperative catalogue a concept similar to ISBD in later years. United efforts were made to bring out uniformity in cataloguing in English speaking countries and Anglo-American Catalogue Code was published in 1908 and in 1949. In the next decades a series of meetings, conferences etc., took place to discuss the standardisation of bibliographic data. Consequently, the various bibliographic standard formats were developed. The outstanding international meetings which were held and themes discussed in them are as follows:

* The International Conference on Cataloguing Principles (ICCP), 1961, Paris. Twelve statements of principles covering standardisation of headings and organisation of alphabetical catalogue.

* The International Meeting of Cataloguing Experts (IMCE)_1969, Copenhagen. It was the beginning of ISBD, the basis of the UBC system.

* The International Symposium on Bibliographic Exchange Formats_ 1978, Taormina,Sicily. Began search for common international exchange format.

Increasing interest in establishing standardised principles for descriptive cataloguing became apparent soon after the ICCP in 1961. Growing use of computers in libraries for electronic data processing and LC's shared cataloguing programme (1966) necessitated the uniformity in bebliographic description for machine manipulation. Development of ISBDs became a core programme of IFLA and a series of ISBDs for various forms of documents were developed and endorsed by international standard organisations.

Information sharing activities among national libraries and bibliographic agencies have demonstrated the requirement of standard format for content designators as well as the format structure for exchange of bibliographic data. Several professional bodies viz., LC, ALA, IFLA, UNESCO, UBC, ANSI, ISO, BSI developed and promoted standard bibliographic formats. The popular standard formats for bibliographic description are ISBDs, AACR2, MARC, UNIMARC, UNISIST Reference Manual, ISO 2709 and CCF.

2.8 GLOSSARY

Content desinator: A code refers to tags, indicators, subfield codes, occurrence identifiers etc. which describe or identify some attribute of a data eelement or group of data elements.

Fixed field: A field which has pre-determined length of characters.

Tag: A tag consists of one or more characters or digits which uniquely denotes a data element or a whole field in a record.

Variable field: A field which can be extended to any length as per requirement.

2.9 REFERENCES AND RECOMMENDED BOOKS

BYRNE, Deborah. *MARC manual: Understanding and using MARC records.* Englewood: Libraries Unlimited. 1991.

CCF: *The Common Communication Format/* edited by Peter Simons and Allen Hopkinson. 2nd ed. Paris: Unesco PGI and UNISIST, 1988.

GREDLEY, Ellen and Hopkinson, Allen. *Exchanging bibliographic data: MARC and other international formats.* Ottawa: Canadian Library Association, 1990.

HARSHA Parekh. "Bibliographic record structure and communication formats". *Lucknow Librarian* 23(1);1991.p.1-24

ISO. Guide for data elements and record format for computer-based bibliographical databases for bibliographic description of different kinds of documents (IS: 11370-1985). New Delhi: Indian Standards Organisation, 1986.

KOKABI, Mortaza. "The internationalisation of MARC: Part I: The emergence and divergence of MARC". *Library Review* 44(4); 1996. p.21-35.

SEN, B. K. "From MARC to CCF". *ILA Bulletin* 27(1), 1991. p.22-37.

VERONA, Eva (1980). *A decade of IFLA's work on the standardisation of bibliographic description.*

2.10 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) What are the compnents in a bibliographic record format?

2) State the main objective of ISBD

3) List out the specialised ISBDs developed for different forms of documents.

4) List out the impotant components in a communication format.

II SHORT NOTES

a) Indian Standard IS:11370-1985

b) UNISIST Reference Manual

c) CCF

Appendix - 1

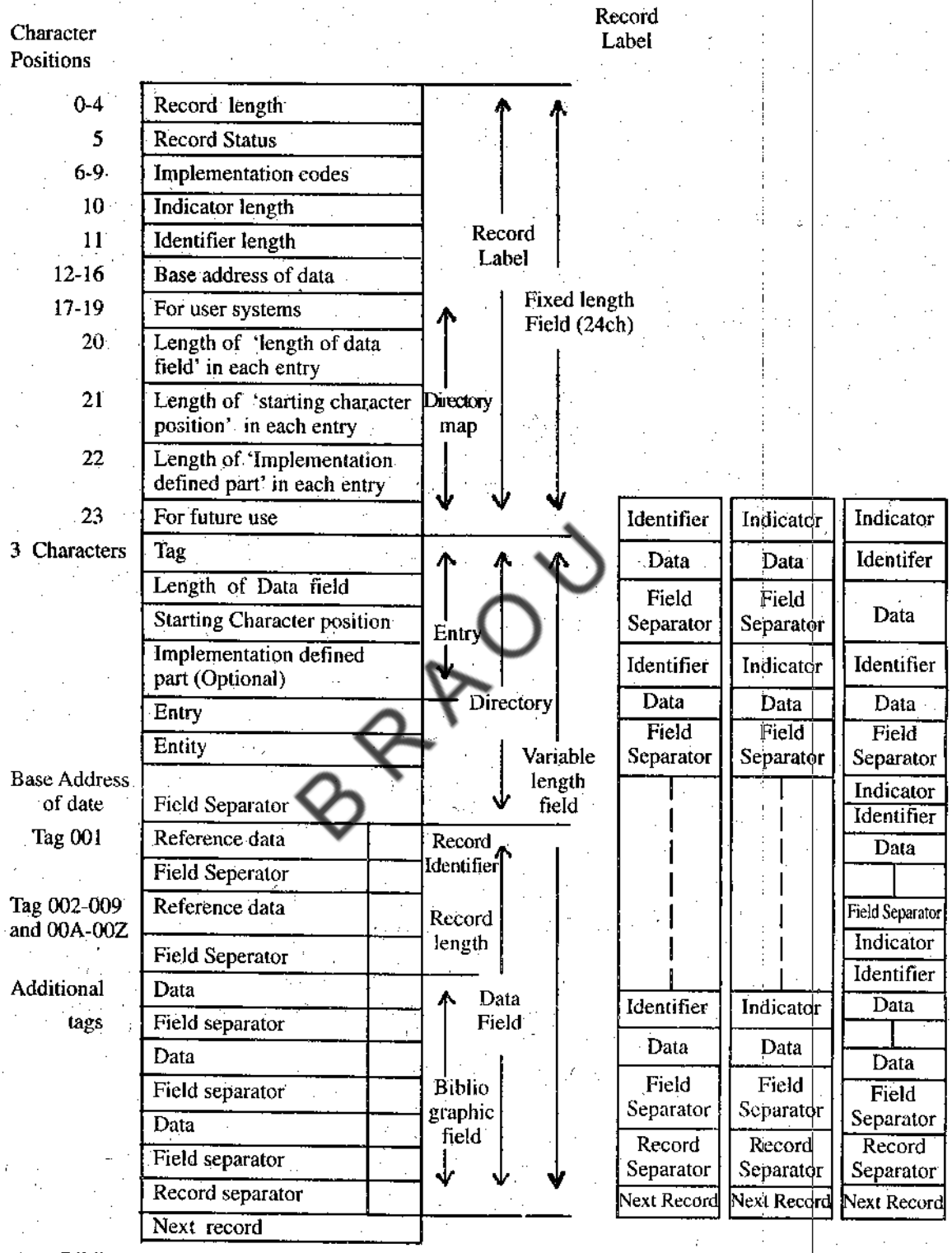
AREAS IN ISBD(G)

Area	Prescribed punctuation (preceding/ enclosing) for elements	Data elements
Note: Each area other than the first, is preceded by a full stop, space, dash, space (. _)		
1 Title and statement of responsibility area		
		1.1 Title proper
	]	1.2 General material designation
	=	1.3 Parallel title
	:	1.4 Other title information
2 Edition area		2.1 Edition Statement
	=	2.2 Paralel Edition Statement
		2.3 Statement of Responsibility relation to the edition
	/	First Statment
	;	Subsequent Statement
	,	2.4 Additional edition Statement
		2.5 Statements of responsibility following an additional edition statement
	/	First Statement
	;	Subsequent statement
3 Material (or type of publication) Specific area		
4 Publication, Distribution area etc		4.1 Place of publication, distribution, etc.
	:	First place
	;	Subsequent place
	:	4.2 Name of publisher, distributor etc.,
	[]	4.3 Statement of function of publisher, distributor etc.,
	,	4.4 Dateof publication, distribution etc.,
	(	4.5 Place of manufacture
	:	4.6 Name of manufacturer
	,)	4.7 Date of manufacture

5	Physical description area		5.1 Specific material designation and extent of item
		:	5.2 Other physical details
		;	5.3 Dimensions of item
		+	5.4 Accompanying material statement
6	Series area		6.1 Title proper of series
	Note: A series statement	=	6.2 Parallel title of series
	is enclosed by parentheses	:	6.3 Other title information of series
	When there are two or more series statements, each is		6.4 Statements of responsibility relating to the series
	enclosed by parentheses	/	First Statement
		;	Subsequent statement
		,	6.5 International standard series number of series
		;	6.6 Numbering within series
		.	6.7 Enumeration and/or title of subseries
		.=	6.8 Parallel title of sub-series
			6.9 Other title information of sub-series
			6.10 Statements of responsibility relating to the sub-series
			First Statement
		;	Subsequent statement
			6.11 International standard serial number of sub-series
		.	6.12 Numbering within sub-series
7	Note area		
8	Standard number		8.1 Standard number (or alternative)
	(or alternative	=	8.2 Key title
	and terms of	:	8.3 Terms of availability and/or price
	availability area	()	8.4 Qualifications (in varying positions)

Appendix-2

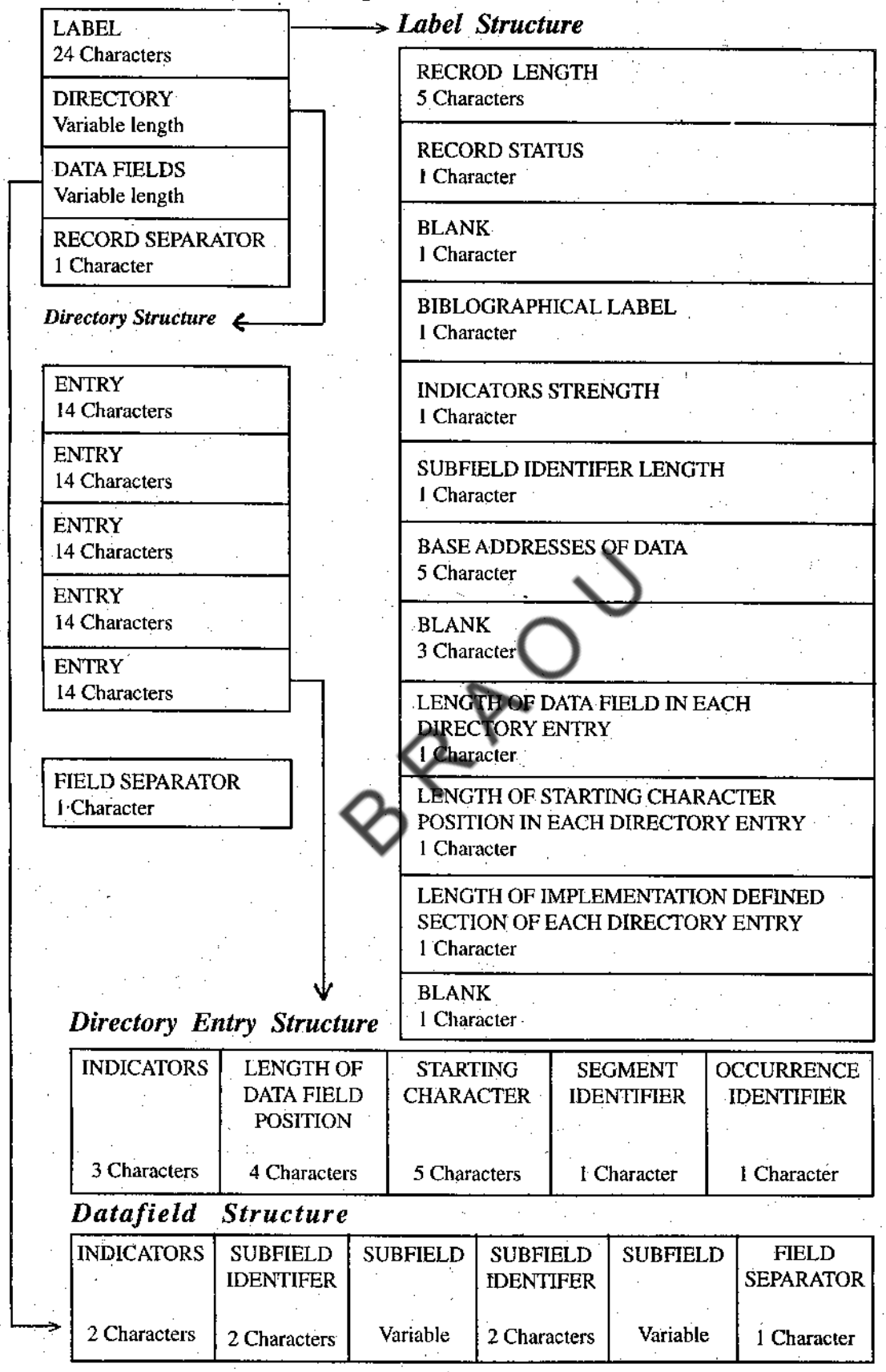
Detailed Record Structure



Bibliographic Field alternatives	(1)	(2)	(3)	(4)
Indicator Length	= 0	= 0	> 0	> 0
Identifier Length	= 0	> 0	= 0	> 0

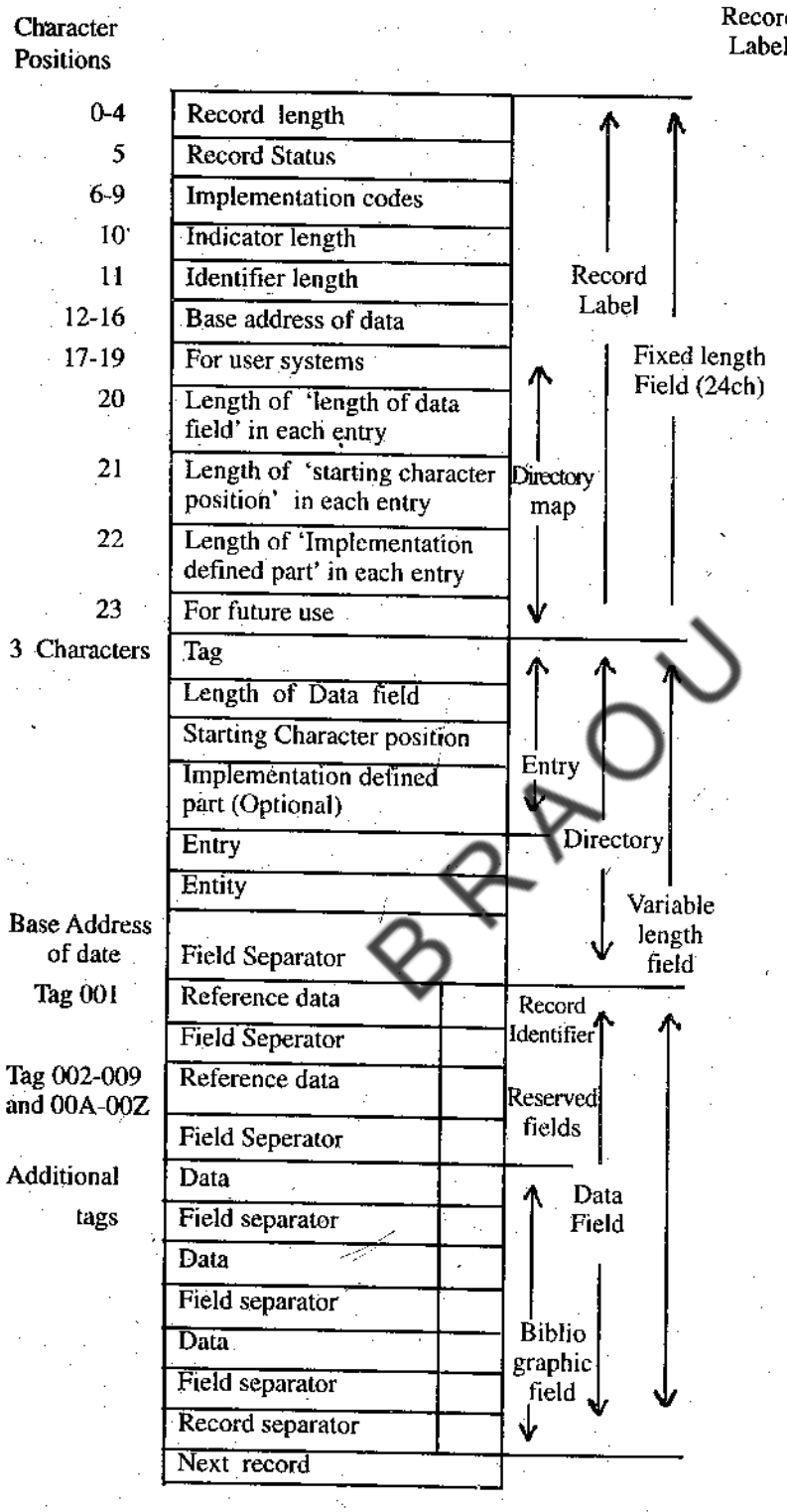
Appendix-3

Diagramatic Representation of CCF Record Structure



Appendix-4

Layout of the Record Format



Body text to show relationships between two bibliographic records and between elements within the same

Appendix- 5

Common Communication Format (CCF)
LIST OF DATA ELEMENTS

TAG	NAME
001	RECORD IDENTIFER
010	RECORD IDENTIFER USED INA SECONDARY SEGMENT
010A	Identifier
011	ALTERNATIVE RECORD IDENTIFIER (R)
011A	Alternative identifier
011B	Identification of agency in coded form
011C	Name of agency
015	BIBLIOGRAPHIC LEVEL OF SECONDARY SEGMENT
015A	Bibliographic level
020	SOURCE OF RECORD
020A	Identification of agency of coded from
020B	Name of agency
020C	Name of code set
020D	Rules for bibliographic description
020L	Language of name of agency
021	COMPLETENESS OF RECORD
021A	Level of completeness code
022	DATE ENTERED ON FILE
022A	Date
023	DATE AND NUMBER OF RECORD VERSION
023A	Version date
023B	Version number
030	CHARACTER SETS USED IN RECORD
030A	Alternative Control Set (C1)
030B	Default Graphic Set (G0)
030C	Second Graphic Set (G1)
030D	Third Graphic Set (G2)
030E	Fourth Graphic Set (G3)
030F	Additional Control Set (R)
030G	Addtional Graphic Set (R)

031	LANGUAGE AND SCRIPT OF RECORD (R)
031A	Language of the record (R)
031B	Script of the Record (R)
041	LANGUAGE AND SCRIPT OF SUMMARY (R)
041A	Language of summary (R)
041B	Script of summary
050	PHYSICAL MEDIUM
051A	Physical medium code(R)
060	TYPE OF MATERIAL
060A	Type of material code(R)
080	SEGMENT LINKING FIELD:GENERAL VERTICAL RELATIONSHIP(R)
080A	Segment relationship code
080B	Segment indicator code
081	SEGMENT LINKING FIELD: VERTICAL RELATIONSHIP FOR MONOGRAPH
081A	Segment relationship code
081B	Segment indicator code
082	SEGMENT LINKING FIELD: VERTICAL RELATIONSHIP FROM MULTI-VOLUME MONOGRAPH
082A	Segment relationship code
082B	Segment indicator code
083	SEGMENT LINKING FIELD: VERTICAL RELATIONSHIP FROM SERIAL
083A	Segment relationship code
083B	Segment indicator code
085	SEGMENT LINKING FIELD:HORIZONTAL RELATIONSHIP(R)
085A	Segment relationship code
085B	Segment indicator code
086	FIELD TO FIELD LINKING(R)
086A	Field linked from
086B	Field relationship code
086C	Field linked to
100	INTERNATIONAL STANDARD BOOK NUMBER(R)
100A	ISBN
100B	Invalid ISBN(R)
100C	Qualification(R)
101	INTERNATIONAL STANDARD SERIAL NUMBER(ISSN)
101A	ISSN
101B	Invalid ISSN(R)
101C	Cancelled ISSN(R)

102	CODEN
102A	Coden
110	NATIONAL BIBLIOGRAPHY NUMBER(R)
110A	National bibliography number
110B	National bibliographic agency code
111	LEGAL DEPOSIT NUMBER(R)
111A	Legal deposit number
111B	Legal deposit agency
120	DOCUMENT IDENTIFICATION NUMBER(R)
120A	Document identification number
120B	Type of number
200	TITLE AND ASSOCIATED STATEMENTS(S) OF RESPONSIBILITY
200A	Title(R)
200B	Statement of responsibility associated with title(R)
200L	Language of title
200S	Script of title
201	KEY TITLE
201A	Key title
201B	Abbreviated key title
201L	Language of key title
201S	Script of key title
210	PARALLEL TITLE AND ASSOCIATED STATEMENTS(S) OF RESPONSIBILITY(R)
210A	Parallel title
210B	Statement of responsibility associated with parallel title(R)
210L	Language of parallel title
210S	Script of parallel title
220	SPINE TITLE(R)
220A	Spine title
220L	Language of spine title
221	COVER TITLE(R)
221A	Cover title
221L	Language of cover title
222	ADDED TITLE PAGE TITLE(R)
222A	Added title page title
222L	Language of added title page title
223	RUNNING TITLE(R)
223A	Running title
223L	Language of running title

230	OTHER TITLE(R)
230A	Other variant title
230L	Language of title
240	UNIFORM TITLE(R)
240A	Uniform title
240B	Number of part(s)(R)
240C	Name of part(s)(R)
240D	Form of subheading(R)
240E	Language of item as part of uniform title(R)
240F	Version
240G	Date of version
240L	Language of uniform title
240Z	Authority number
260	EDITION STATEMENT AND ASSOCIATED STATEMENT(S) OF RESPONSIBILITY(R)
260A	Edition statement
260B	Statement of responsibility associated with edition(R)
260L	Language of edition statement
300	NAME OF PERSON(R)
300A	Entry element
300B	Other name elements
300C	Additional elements to name
300D	Date(s)
300E	Role(coded)(R)
300F	Role(non-coded)(R)
300Z	Authority number
310	NAME OF CORPORATE BODY(R)
310A	Entry element
310B	Other parts of name(R)
310C	Qualifier(R)
310D	Address of corporate body
310E	Country of corporate body
310F	Role(coded)(R)
310G	Role(Non-coded)(R)
310L	Language of entry element
310S	Script of entry element
310Z	Authority number
320	NAME OF MEETING(R)
320A	Entry element
320B	Other parts of name(R)
320C	Qualifier(R)
320E	Country
320G	Location of meeting

320H Date of meeting(in ISO format)
320I Date of meeting(in free format)
320J Number of meeting
320L Language of entry element
320S Script of entry element
320Z Authority number

330 AFFILIATION(R)
330A Entry element
330B Other parts of name(R)
330C Qualifier(R)
330D Address(R)
330E Country of affiliation
330L Language of entry element

400 PLACE OF PULICATION AND PUBLISHER(R)
400A Place of publication(R)
400B Name of publisher
400C Full address of publisher(R)
400D Country of publisher(R)

410 PLACE OF MANUFACTURE AND NAME OF MANUFACTURE(R)
410A Place of manufacture(R)
410B Name of manufacture
410C Full address of manufacture(R)
410D Country of manufacture(R)

420 PLACE AND NAME OF DISTRIBUTOR(R)
420A Place of distributor(R)
420B Name of distributor
420C Full address of distributor(R)
420D Country of distributor(R)

440 DATE OF PUBLICATION(R)
440A Date of formalized form
440B Date of non-formalized form

441 DATE OF LEGAL DEPOSIT
441A Date of legal deposit

450 SERIAL NUMBERING
450A Serial numbering and date

460 PHYSICAL DESCRIPTION
460A Number of pieces and designation
460B Other descriptive details
460C Dimensions
460D Accompanying material(R)

465	PRICE AND BINDING(R)
465A	Price(R)
465B	Binding(R)
465C	Date of price(R)
480	SERIES STATEMENT AND ASSOCIATED STATEMENT(S)
481	OF RESPONSIBILITY(R)
480A	Series statement
480B	Statement of responsibility associated with series statement(R)
480C	Part statement
480D	ISSN
480L	Language of title
480S	Script of title
490	PART STATEMENT(R)
490A	Volume/part numerator and designation(R)
490B	Pagination defining a part
490C	Other identifying data defining a part
500	NOTE(R)
500A	Note
510	NOTE ON BIBLIOGRAPHICAL RELATIONSHIP(R)
510A	Note
520	SERIAL FREQUENCY NOTE(R)
520A	Frequency
520B	Dates of frequency
530	CONTENTS NOTE(R)
530A	Note
600	ABSTRACT(R)
600A	Abstract
600L	Language of abstract
610	CLASSIFICATION SCHEME NOTATION(R)
610A	Notation(R)
610B	Identification of classification scheme
620	SUBJECT DESCRIPTOR(R)
620A	Subject descriptor
620B	Identification of subject system

R=Repeatable field

UNIT - 3 : BIBLIOGRAPHIC DESCRIPTION FOR NON-PRINT MATERIALS

Structure

3.0 AIMS AND OBJECTIVES

In the previous unit we have discussed about the standards for bibliographic description. The present unit discusses the bibliographic description of non-print materials.

After reading this unit ,you will be able to

- describe the role of non-print materials as sources of information
- identify varieties of non-print materials
- list out the standards for bibliographic description of non-print materials
- provide bibliographic descriptions to non-print materials according to AACR-2R.

3.1 INTRODUCTION

The story of written communication dated back to the times when nomadic people communicated through cave drawings. In ancient times people used clay tablets, papyrus rolls, parchment rolls, palm leaves etc., as writing media for communication. A break-through occurred in the 15th century when Johan Gutenberg invented printing from the movable types. Further developments in printing technology took place during the great Industrial revolution in the West. In the next five centuries printed documents dominated the world as the most convenient medium for communication. The world is now witnessing exponential growth of information and there is a need to provide quick access to information. The exhorbitant cost of printing and publishing, the quickly outdated feature of print media and the need to reduce time lag in conventional printing necessitated the shift in focus from conventional print media to non-print media. Over the last two decades, there has been tremendous growth in the development of new technologies in photographic, micrographic, computers, fibre optics and telecommunication technologies that affect the preparation, organization, storage and retrieval of information.

Presently, there is a trend towards compressing information carriers so that they occupy less space and make storage and distribution easier. Information seekers emphasize on directly accessing information stores/databases. Compression of information led to the growth of microforms. Optical fibre technology will enhance the ability of the people to communicate directly with information source using network computer terminals.

Non-print materials (NPMs) as storage media have been gaining importance since the middle of the twentieth century in providing potential alternative access to information. NPMs have enormous information storage capacity with low cost. Data damage is drastically reduced with NPMs and information can be retrieved and transferred speedily and accurately. NPMs with their durability feature provide the benefit of repetitive use of information without deterioration or loss of information. In view of certain advantages users prefer to consult materials in non-print format. It is, therefore, important you should realise the potential features of various non-print formats as effective storage, information retrieval and dissemination media.

3.2 NON-PRINT MATERIALS

The term "Non-print materials" also used synonymously with the terms "Audio-visual materials" and "Non-book materials" which refers to any format not in book or similar print format. Formats other than print formats include a multitude of formats from realia (objects) to most sophisticated forms of silicon technology to fibre optics technology. The formats under consideration in this unit, however, are those formats which are most likely to be found in libraries and information units and which require special treatment in terms of their bibliographic description in order to exploit information from those formats. The important NPM formats are described below.

3.2.1 Microforms

Microforms are miniature reproductions of printed or other graphic matter, which cannot be utilized without magnification by specific equipment. Microfilms, Microfiche, Micro-opaque and Aperture cards are included under the heading and designated as "microforms".

i) ***Microfilm:*** It is stored on a roll or reels of film available in 35mm or 16mm format. Microfilms may be with or without sprocket holes, with or without self-threading cartridges and may be in black and white or in colour. Microfilms can hold large quantities of information in a very little space at very low cost.

ii) ***Microfiche:*** In a microfiche, micro images are arranged in grid formation on a transparent film of 3x5 inches or 4x6 inches either in colour or black and white. A strip of eye-reading is placed along the top edge for easy identification of the material. Ninety-eight frame format is more common for microfiche of documents using a reduction ratio of 24x from the original.

iii) ***Micro-opaque:*** The micro images are produced on a card rather than on a film, either photographically or by an offset litho press. As micro-opaques are relatively inefficient and they have become obsolete formats.

iv) ***Aperture cards:*** These are pieces of card with a window into which the microfilm is inserted. Aperture card is approximately 7.75 x 3.25 inches. It can be written on for reference purposes and notched for mechanical sorting for computer information retrieval.

3.2.2 Sound recordings

A recording on which sound vibrations have been registered by mechanical or electronic means so that the sound may be reproduced. There are many formats of sound recordings

Audio Tapes : Audio tapes or sound tapes are found in two formats - Open reel and Cassette. These formats each require its own playback equipment because of speed and tape width variation.

i) ***Open reel or Reel-to-reel tape:*** The usual width of the tape is 1/4" and being wound on one reel and then fed to another open reel. This format is used mainly for in-house production of masters or originals as the quality of Open reel tape ensures a higher standard of recording and at a faster recording speed. The reels vary in size usually 8 cm, 13 cm, 18 cm or 26.5 cm in diameter. Playing time depends on the speed at which the recording is done and the thickness of the tape.

ii) ***Cassette tape:*** Audio cassettes have a standard dimensions of 10.2 x 6.4 cm with a standard tape speed of 4.75 cm/s. C30, C60, C90 and C120 cassettes are available at standard timings of 15 minutes, 30 minutes, 45 minutes and 60 minutes respectively of each side. Cassette tapes for computers at standard timings are also available.

3.2.3 Visual Formats

i) ***Transparencies:*** Transparencies are extensively used in education and training environments. Transparencies are made from plastic in sheet or roll formats. Visuals can be hand-written or drawn with specific quick drying ink pen. A special overhead projector is used for projecting the visual on to a screen.

ii) ***Film:*** Photographic film is a polyester-based material coated with a layer of emulsion. Images are created as a result of a chemical reaction to light and further fixed and chemically washed. There are a number of audio-visual formats based on film and each format requires special equipment to exploit the material.

a) ***Cinefilm:*** A Cinefilm is a sequence of images arranged vertically and as the film passes through the projector, the illusion of movement is created on to a screen. A variety of Cine film formats is seen. You will find usual 35mm or 70mm cinefilm with sprocket holes on both sides of the tape. Besides this, 16mm film with optical sound track which cannot be erased unlike magnetic sound track. Sprocket holes are found on one side of the tape only, the other side is earmarked for sound track. The film is projected at 25 frames per second. 16mm silent films with sprocket holes on both sides also available and can be projected at 1 frame per second.

b) ***Filmstrip:*** The filmstrip is a collection of images organized in two different forms on a continuous piece of film. A short length of such a film, sometimes mounted in a rigid format is called a Filmstrip. There are two kinds of filmstrip. One being single (half) frame and the other double (full) frame. The single format is arranged vertically, while the double format is arranged horizontally on the strip. Suitable projectors are required for these two formats for operation.

c) ***Slides:*** Slides are individual frames of film, mounted on storage cases. The mounts can be cardboard or plastic or glass. Slides are placed in a slide box with sets serially numbered. Two common formats viz., 35mm and 110mm format. Slides are projected on to a screen with a slide projector.

3.2.4 Videograms

Videograms are programmes of moving picture and sound which has been recorded magnetically on to a tape or electronically put on to a disc. Unlike audiotapes, there is no standardization in the production of videotapes and as such they are mostly dependent on specific kind of video play back unit.

i) *Open reel videotape:* The use of open reel tape is very limited and mainly used for broadcast purpose and production of master material only.

ii) *Videocassettes:* This is the most popular video format in use containing 1/2 and 3/4 inch. You find a number of formats with varying sizes and varying recording timings. Care is necessary regarding compatibility of tape, recording system and playback equipment.

iii) *Video cartridge:* This is an open reel with container. During operation the tape is automatically threaded and played.

iv) *Videodisc:* A laser beam scans the disc and the image appears on the visual display unit of a computer or television screen. A 12" Videodisc can contain 30 minutes of motion picture with two independent sound tracks or 5400 separate colour still frame images with no sound. The vast storage capacity of Videodisc with fast retrieval of information is most common in libraries.

3.2.5 Magnetic Discs

Magnetic discs are used with computers for recording and retrieving information. Information can be written on to a magnetic disc, which is in machine readable form. Magnetic disc drives are of two types: fixed disc unit and removable disc unit.

a) *Floppy disc:* The Floppy disc or Flexible disc replaced punched cards used earlier for data entry. These discs are made of plastic and coated with magnetic material. The popular sizes of the disc are 8", 5.25" and 3.5" in diameter.

b) *Hard disc:* Hard or fixed discs are available for all configurations of computers and can be permanently installed in a computer or removable cartridges or disc packs.

3.2.6 Optical Data Storage Media

Optical technology for storage of information and retrieval involves the use of laser beams. The most exciting developments in this technology in recent years have been the creation of new types storage media used in computers such as optical laser discs, optical card and optical tape. Optical storage media offer convenient data handling of vast information storage and retrieval. Laser optical discs are hard metal discs ranging in size from 4.72" to 14". Most optical discs are read only devices.

i) *CD-Audio:* This compact disc is a popular format for distributing music in disc form. The standard size of the disc is 12 cm in diameter and 1.2 mm thickness and have a central hole of 15mm diameter. Maximum playing time of one disc is approximately one hour.

ii) *CD-ROM:* A common version of laser disc is the CD-ROM. CD-ROM stands for Compact Disc Read only Memory. The potential of CD-ROM lies in its compactness, 12cm in diameter and 1.2mm thickness, portability, reduced shelf space and its low maintenance cost. A single disc supports upto 660 Mega bytes data i.e. approximately equivalent to 470 numbers of 1.44 MB floppy discs; up to 18 hours sound; up to 74 minutes of movie or video. Like audio CD, a CD-ROM disc physically consists of a metallic disc branded to a polycorbonate-base material. This is coated with a transparent protective layer. With its read only feature, CD-ROM is immensely useful as a publishing and distribution medium. CD-ROM offers unequalled advantages in libraries and information centers in terms of large data storage format for information retrieval and dissemination.

iii) *High Density CD-ROM:* These discs could hold full-length movies, high quality music and complex computer data. The new format is named as Digital Video Disc (DVD). The new format looks as like CD but is double sided. Each side with 4.75 GB of space that can hold 133 minutes of video or equivalent of eight CDs of music.

iv) *CD-I:* Interactive CD carries combination of text, graphics, audio, stills and moving pictures.

v) *CD-V:* Video CD carries audio in digital recording and video in analog. The general size is 12cm, carrying up to 6 minutes of video with sound and further 20cm carrying up to 20 minutes of audio and video on each of the two sides. 30cm carrying up to 60 minutes of audio and video on each of the two sides.

3.3 STANDARDS FOR DESCRIPTION OF NON-PRINT MATERIALS

Non-Print Materials are mostly dependent on equipment such as microform readers, video recorders, tape recorders players, projectors (slide, film, etc.), overhead projectors, computers, disc drives etc. As you see information is hidden in NPMs and has to be seen or heard through above said media. Unlike NPMs, printed media (i.e. book) is ready for immediate reading. As you cannot browse NPMs like books on a shelf, it is necessary to describe the contents of NPM in full and in greater specificity. While describing NPMs focus should be for accessibility of information on NPMs. One must have the knowledge of the format of NPM to be catalogued. It is often difficult to identify and obtain bibliographic elements from NPMs for standard bibliographic description.

There have been concerted attempts to provide standard rules for bibliographic description for NPMs since the 1950s. Library of Congress had published its Supplementary Rules Descriptive Cataloguing in Library of Congress: Phonorecards in 1952; Motion pictures and Filmstrips in 1953; Pictures, designs and other two dimensional representations in 1959. Canadian Manual for Non-Book Materials originally published by Canadian Library Association in 1970 was later updated in 1979. The British Library Association Media Cataloguing Rules Committee published "Non-Book materials cataloguing rules: Integrated code for practice and draft revision of British Text Part III" in 1973. This is Popularly known as LANCET rules designed to be used in conjunction with AACR. IFLA's standard bibliographic description ISBD (NBM) appeared in 1977. Detailed rules for description of NPMs are given in AACR II R and there is a growing trend towards the use of this code for NPMs organisation.

3.3.1 ISBD (NBM) Framework

ISBD (NBM) provides the following structure for bibliographic description.

1. Title and statement of responsibility area

Title statement : Title is given by the author/producer/publisher of NPM. Title can be identified as follows:

Uniform title : It is a particular title by which a work that has appeared under varying titles is to be identified for cataloguing purposes.

Supplied titles: You often come across NBMs have no title and one must be supplied. Such a situation is very common with illustrations, specimens and models.

Collective title: An individual item may contain several works. There may be two or more titles associated with its description. If there is not a collective title associated with the work then it will be necessary to record each title in the order in which they appear in the chief source of information.

2. ***Responsibility for the creation of NBM :*** There might be several collaborators such as performers, adapters, producer, narrator etc. who may have shared responsibility in creating the NBM. These collaborators also form part of description.

3. ***Physical description of materials:*** This includes the extent of the item, other physical details, dimensions and accompanying material.

3.3.2 AACR-2R

Rules for description are elaborately given in AACR (second edition 1988 Revision) generally referred to as AACR 2R. The broad framework for description of all materials is given in the general chapter 1 in AACR 2R. While following the rules laid down in Chapter 1 General rules for description it is important to keep in mind the following points.

1) The physical description of any item should be based in the first instance on that chapter in AACR 2R dealing with the class of materials to which that item belongs.

2) It may be necessary that only national bibliographic agencies will adopt "Third level of description" for recording all the bibliographic elements pertaining to an item to be catalogued and all other agencies may choose the "first or second level of description".

3) The bibliographic description prescribed will not normally be used by itself, but will usually form part of a complete entry in a catalogue or other bibliography. The headings, classification numbers etc. used in arranging entries in a catalogue do not form part of the standard description of an item. The framework of chapter 1 in AACR 2R will give the physical description as outlined below:

1.1 Title and statement of responsibility

B. Title proper
C. General Material Designation [ ]
E. Other title information :
F. Statement of responsibility /
Subsequent statement of responsibility ;

1.2 Edition area

B. Edition statement . —

1.3 Material (or type of publication) specific details area . —
(No specific use of this area for NBM except file characteristic details for computer files)

1.4 Publication, distribution , etc. area

B. Place of publication, distribution etc. . —
D. Name of the publisher, distributor etc. ;
E. Statement of function of publisher, distributor etc. :
F. Date of publication, distribution etc. ,
G. Place of manufacture, name of manufacture, date of manufacture;

1.5 Physical distribution area

B. Extent of item	. —
C. Other physical details	:
D. Dimensions	;
E. Accompanying material	+

1.6 Series area

B. Title proper of series	. — (...)
G. Numbering within series	;

1.7 Note area

B. Notes	. —

1.8 Standard number and terms of availability area

B. Standard number (or alternative)	. —
D. Terms availability	:

The numbering of the above framework refers to the specific AACR-2R rules as you find 1) All the elements that are required to describe NPM 2) Assigns order to these elements; and 3) prescribes punctuation for these elements.

Authorship is generally diffused and cannot be established as readily for many non-print materials. As such entry under title will occur more frequently for non-print materials. The following rules of main entry are applied in the order in which they are listed.

1) A reproduction of a work originally produced in another medium is entered in the same manner as the original work

2) A work for which authorship can be clearly established is entered under the author. Authorship is not normally attributed to consultants, performers, producers, directors, designers etc., except in case of a sound recording, motion picture, or video recording in which the responsibility of the performing group goes beyond that of performance is entered under the performing group.

3) An item without a collective title is entered under the heading appropriate to the first work.

4) A work for which authorship cannot be established is entered under title.

3.4 SOURCES OF BIBLIOGRAPHIC ELEMENTS FOR THE DESCRIPTION

The chief source of information for bibliographic description of printed monographs is the title page. Such a prime source for the description of NPMs is unusual as mostly information being scattered around the NPM document and its container. AACR 2R used the

concept of the "Chief source of information" in relation to specific NPM to overcome this problem. Thus, you will find the chief source of information under the following categories:

1) The material itself, including the container, where this forms part of the item. E.g. Cassette or Cartridge.

2) The container itself where this is completely separated from the item, for example a box.
3) Accompanying material i.e. printed text guides and other leaflets issued with an item.

4) Other reference works/catalogues etc.

The order of preference for each of the above categories is given in AACR 2R by the specific material chapter.

3.5 BIBLIOGRAPHIC DESCRIPTION OF NPMs

i) ***Sound recordings:*** Sound recordings of various kinds viz., disc, tape (open reel-to-reel), cassette, cartridge, roll, sound track film lack standardization in presenting information about themselves. The content of recording, the person (s), who has made the recording, technical details of recording and the equipment requirements for its use may be found on the disc/tape/reel/cartridge and the label. This has to be regarded as the chief source of information. But you may need to be sought information from accompanying textual material such as a booklet or record sleeve or box.

Chapter 6 in AACR 2 (Second edition Revision 1988) details various bibliographic elements to be shown and their order in the catalogue entry. Besides giving the standard description, the physical description area are to be shown as set out under various rules: the number and kind of physical units, playing time, type of recording i.e. the way in which the sound is recorded on the item (analog or digital for disc or cassette and optical or magnetic for sound track film reel), playing speed for analog disc in revolutions per minute(rpm) and digital disc in metres per second(m. per sec.), analog tape in inches per second (ips), sound track film in frames per second (fps), track configuration for sound track film (centre track or edge track), number of tracks and number sound characters (mono, stereo, quad). Dimensions of sound recordings are given in inches. You may have to make elaborate notes about the nature of item being catalogued, the performers and the recording details.

Workedout example for a Sound Recording

Fry, D. B.

Science looks at speech [sound recording] / D. B. Fry;
Interviewer Paul Vaughan. — London: Seminar Cassettes,
[1977].

1 sound cassette (51 min.) : 1 7/8 ips, mono. —
(University series)

Seminar cassettes: SS105

ii) ***Motion Pictures and Video Recordings:*** Rules are set out in chapter 7 of AACR 2R for description of motion pictures of all kinds including complete films and programmes, compilations, trailers, newscasts and news films, stock shots and unedited material. The chief source of information is the title frame or the container label. The cataloguer may have to supply titles for commercial advertising film, which includes product or service and the word "advertisement". Supplied titles for unedited material and news films include place, date of event, personalities and subjects. Parallel titles are common because of the world-wide distribution of films. The statement of responsibility will name producer, director or animator. The physical description is peculiar to the medium. You may have to record appropriate terms such as film cartridge, film cassette, film loop, film reel, video cartridge, videocassette, and video disc and video reel for specifying the physical units. The playing time, sound characteristics i.e. "sd" for sound, "si" for silent, colour, other technical characteristics like projection speed in frames per second and playing speed in rpm which affect showing are also to recorded. Notes may be necessary to record the names of actors, players, performers and/or presenters and others concerned in making a film. Cataloguer can make elaborate notes for additional physical details on special sound characteristics, length of film, colour recording system, print form, video recording system etc. Access points will often be titles. Other access points may be based on the names of any of the persons or corporate bodies to whom responsibility for a film has been attributed.

Workedout Example for a Motion Picture

Aids for teaching the mentally retarded [motion picture] / Production James B. Henderson, Spencer Nelson and Oakleigh Thorne II. — Boulder, Colo.: Thorne Films; Ipswich: Concord Films [distributor], 1964.

5 film reels (ca 40 min.): sd., col.: 16 mm.

One phase reel.

iii) ***Graphic Materials:*** These are flat and static materials. These materials consist of opaque two dimensional art originals, reproductions, charts, photographs, technical drawings or projectable or viewable material for example film strips, radiographs and slide transparencies. The terms to be used to designate the specific media are listed at 8.5B1 in AACR 2R. The words cartridge or reel are to be added to filmstrip or steriograph when appropriate. The number of frames or double frames are added to the designation of filmslip or filmstrip while the number of pairs of frames are added to the designation of steriograph. The number/approximate number of sheets are added to to the term flipchart and number of overlays or attached overlays if any are added to the term transparency. Medium specific other physical details are to be given for example medium (chalk, oil, pastel) and base (board, canvas, fabric) for art originals; process in general terms (engraving, lithograph) or process in specific terms (copper engraving, chromolithograph) for art prints; double or single sided for charts and flip charts; sound

characteristic for filmstrips, filmslip, flash cards and slides if sound is integral; methods of reproduction (blue print, photocopy) for technical drawings. Colour details (e.g. col. or b&w, sepia) have to recorded for all graphic media except art originals, radiographs and technical drawings. Dimensions of height x width in centimeters will have to be given for all graphic materials except filmstrips, filmslip and stereograph while only width (gauge) of the film is given in millimeters for filmstrips and filmslip.

Workedout Example for a Filmstrip

Newton, Stella Mary

Napoleon [filmstrip] / Compiled and annotated by Stella Mary Newton. — London: Visual Publications, 1971.

2 filmstrips: col. ; 35mm. + 2 manuals (32p.;27cm.). — (A closer look into history; 6)

Contents: 1. The man and his surroundings (40 fr.) — 2. His effect on his time (37 fr.)

iv) ***Microforms :*** Microforms are quite useful for reproduction of voluminous works to a few cards or reels of films. Rare and fragile materials could be reproduced and make them transportable in reduced facsimile. Though microforms are reproductions they are to be catalogued as microforms only. The title proper, statement of responsibility, edition, imprint etc., are to be taken from their own title frame or title card or wherever the fullest information is given (11.0B of AACR 2R). The physical description must state the medium as appropriate i.e. aperture card, microfilm, microfiche or micro-opaque together with the number of physical units in Arabic numerals and its dimensions. In addition to the elucidatory notes as set out in rules that may be required for any item to be catalogued, microforms need a note on the reduction ratio, if the item is not conformed to the standard reduction ratio of 16x-30x range, and also the name of the reader on which microform cassette or cartridge which effect the use of it. You will also record the details of original works for those microforms that reproduces the existing works. Cataloguer has to provide a uniform title in order to collocate the heading for the microform with the heading for the original work or to act as its substitute.

Workedout Example for a Microform

Croghan, Antony.

A thesaurus-classification for the physical forms of non-book media [microform] / by Antony Croghan. — 2nd ed. — London: Coburgh publications, 1976.

1 microfiche (53 fr.): negative; 11 x 15 cm.
First ed. published in book format.

v) ***Three-Dimensional Artefacts and Realia :*** Chapter 10 of AACR-2R covers the description of three dimensional objects that occur in nature or prepared by individual. Rules for description are set out for naturally occurring objects, generally prepared as exhibits or specimen, together with microscope slides holding such specimen, original art in three dimensions and reproduction of them, machines, clothing, furniture etc. The medium Braille cassette, which is of interest in special education libraries, is also included in this chapter. In addition to the said media, the covers commercially produced material for general distribution like games and puzzles. These media display sufficient information for identification and description. If the material has no label or authenticated origin, the cataloguer must supply title and description about the object according to the guidance of the code. All the elements in respect of physical units, other physical details such as material of the medium of which it is made of, colour, dimension etc. have to be recorded as set out in rules.

Workedout Example for a Model

Human brain kit [model]. — Skokie, Ill.: Lindberg, 1974.

1 model (4 pieces); plastic; in box, 36 x 26 x 6 cm. +
1 guide (4p. : ill.; 28cm.). — (Natural science series; no. 1306)

Container title

vi) ***Computer Files :*** Advances in computer technology, particularly the development of microcomputers, have led to greatly increased production and use of computer files. Libraries have to maintain lot of data in the form of computer files and programmes. Based on the ISBD (CF) and other standards available chapter 9 of AACR2 has been re-written. The chief source of information is the title screen. File characteristics data, which includes computer data or computer programmes and the number of records or bytes, are to be recorded. The physical description includes the terms - computer tape cartridge, computer disc or computer optical disc. System requirements precede all other notes except any necessary note on the nature and scope of the file (9.7B1 of AACR 2R).

Workedout Example for a Computer File

Annual review of phytopathology. 1986-1995 [computer file].

Computer data. — Pal Alto, CA.: Annual Reviews, 1996.

1 computer laser optical disc; 4 3/4 in.
Project editor: Gregory Shaner.
Financial support: Kenneth F. Baker

System requirements: IBM PC 486, 8MB RAM;
Compatible MS-DOS 3.1 or later and Microsoft
Windows 3.1 or later in enhanced mode.

3.6 LET US SUM UP

The developments in new technologies have offered tremendous scope for vast storage of information coupled with quick access to information source. Efficient retrieval and transfer of information without loss are possible with the repetitive use of non-print materials. The use of new technologies has become *sine qua non* in libraries and information units to cope up with information explosion of which the world is witnessing. Improvements in technology are leading to the creation of information access oriented mechanisms in the form of new non-print media and the libraries are transforming to paperless information systems. Librarians are key players to provide catalogues, databases and other access devices, which would provide scope for productive use of information resources. The bibliographic description of non-print materials is more or less same as that of print media with exception to physical description of non-print materials. The format is unique with each non-print media. AACR 2R could be used as an effective standard for bibliographic description for non-print materials in libraries and information units.

3.7 GLOSSARY

Chart: A sheet of information arranged in tabular or graphic form produced on an opaque backing.

Chief source of information : The source of bibliographic data to be given first preference as the source from which information for a bibliographic record is taken.

Diorama : A scene produced in three dimensions by placing objects, figures etc., in front of a representational background.

Game: A set of materials designed for play according to set of rules.

Kit : A set of material composed of many textual parts, or two or more media, none of which is identifiable as the predominant constituent of the item.

Microscope slide : A specialized slide produced specifically for use with a microscope or microprojector.

Model : A three dimensional representation of an object, either exact or to scale, a mock-up.

Overlay : A transparent sheet designed to be superimposed on another sheet for modification of the original data.

Papyrus Rolls : Rolls of writing material made of thick fibrous stems of papyrus water plant. Early Egyptians used papyrus rolls for writing purposes. The word paper is derived from papyrus.

Parchment Rolls : Writing material made of thin layers of animal skin.

Technical drawings : A plan, elevation, cross section, detail, diagram, perspective etc. made for use in an engineering, architectural or other technical context.

3.8 REFERENCES AND FURTHER READING

FLEISCHER, Eugene B. and Goodman, Helen. *Cataloguing of audio-visual materials.* New York: Nea-Schuman, 1980.

FLEISCHER, Eugene B. and Goodman, Helen. *Cataloguing of non-print media: a manual based on the Anglo-American Cataloguing Rules II.* New York: Nea-Schuman, 1980.

FOTHERGILL, Richard and Butchart, Ian. *Non-book materials in libraries.* 2nd ed. London: Clive Bingley, 1984.

FROST, Carolyn O. *Cataloguing non-book materials: problems in theory and practice.* Littleton, Colo.: Libraries Unlimited, 1983.

HUNTER, Eric J. and Fox, Nicholas. *Examples illustrating AACR 2.* London: The Library Association, 1980.

IFLA. I*SBD NBM): International Standard Bibliographic Description for Non-book materials.* London: IFLA, 1977.

Olson, Nancy B. *Cataloguing of audio-visual materials: a manual based on AACR* 2. Mankato, Minn.: Minnesota Scholarly Press, 1981.

Weihs, J. R. *Non-book materials: the organisation of integrated collections.* 2nd ed. Ottawa: Canadian Library Association, 1979.

3.9 ASSIGNMENT

Prepare entries for the non-print materials available in your library as per AACR2 rules. If non-print materials are not available in your library, go to a university library or special library or information centre and collect a sample of ten different formats and prepare the entries.

3.10 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) List out the varieties of NPM and their advantages over print media

2) Identify the standard areas of bibliographic description for NPMs.

3) Give examples of physical description of the following:

   i) One sound disc
   ii) One Video cassette
   iii) One transparency with five overlays
   iv) One Computer disc
   v) One microfilm reel
   vi) One sound track film reel

II SHORT NOTES

a) AACR-2R

b) ISBD (NBM) Frameork

BLOCK - II : SUBJECT ANALYSIS AND INDEXING

Though there are many approaches that are helpful in locating documents in libraries, subject approach forms the scientific basis for the classification and arrangement. The basic purpose of any classification is to individualise each subject and then assign the subject a number or a term as its content identifier. Classification schemes follow the first pattern for a dual purpose of content identification and shelf arrangement in libraries, while the indexing systems use the latter one as document surrogates in information retrieval systems.

A General Classification scheme embraces the entire domain of the universe of knowledge, while special scheme aims at a narrow domain of knowledge, a particular user group, or caters for a form of material, which requires special treatment. In the first unit (Unit-4) of this block, we briefly discussed the general classification schemes (Dewey Decimal Classification, Library of Congress Classification, Colon Classification and Bibliographic Classification) and Special Classification schemes (Moy's Classification, INSPEC Classification, London Classification of Business Studies). The second unit (Unit-5) is exclusively devoted to Universal Decimal Classification (UDC).

Subject Indexing is a method of information retrieval. It is based on the conceptual analysis of the subject of documents. Indexing operations, which were performed by the human indexers for quite a long time, now being developed as automated systems replacing human experts partially or completely. There are three units devoted to indexing systems in this block. An overview of Subject Indexing (Unit-6), extensive discussion on PRECIS and POPSI and their comparison (Unit-7), and Thesaurus- Its structure, functions and construction (Unit-8) have been included as essential components of this block.

In this block, you find the following five units:

Unit-4: Classification Schemes - General and Special

Unit-5: Universal Decimal Classification (UDC)

Unit-6: Subject Indexing - An Overview

Unit-7: PRECIS and POPSI

Unit-8: Thesaurus - Its Structure, Functions and Construction

UNIT - 4 : CLASSIFICATION SYSTEMS - GENERAL AND SPECIAL

Structure

4.0 AIMS AND OBJECTIVES

Classification, more than anything else, has given librarianship the status of science. In this unit we trace the evolution of general schemes of classification which attempt to cater for all knowledge, to the present. The main emphasis is on those schemes which have been well established and continue to retain a practical function in a library.

After studying this unit, you will be able to

- describe the major contributions to library classification
- list out the principal features of general and special schemes of classification with suitable examples.

4.1 INTRODUCTION

Earlier attempts at classification were really attempts to organise human thought, as they were designed to aid the mental plotting of the universe of thought and objectives, rather than systems. A general classification embraces all knowledge proceedings for the classification of knowledge recorded in documents. Most of the general schemes of classification came into existence. A classification whose area of purview extends to the entire domain of the universe of knowledge, and for classifying documents in all media is known as General Classification System.

The earliest recorded scheme was that designed by 'Callimachus' for the library of the Pharaohs at Alexandria (260-290 BC). The era of modern classification started in 1876 with the publication of Dewey Decimal Classification (DDC). Since then, a number of general classification systems have come up to organise the knowledge. Some of the other general classification systems are -

1) Expansive Classification (EC), 1891-1893, designed by Charles Ammi Cutter

2) Subject Classification (1906), designed by the influential British Librarian, J.D. Brown.

3) International Classification (1961), by another distinguished American Librarian, Fremont Rider.

4.2 THE FIVE MAJOR CLASSIFICATION SCHEMES

There are some more general schemes made and used locally, though there are eight general systems of classification in their historical setting. Let us discuss the following five major general classification systems, that are in use today.

4.2.1 The Dewey Decimal Classification (DDC)

The first modest edition of Dewey Decimal Classification, consisting a total of only 42 pages, appeared in 1876. In 1989 the 20th edition was published in four volumes. Volume-1: The Introduction and Auxiliary Tables; Volume-2 and 3: The Schedules; and Volume-4: The Index with - for the first time a manual of practice. DC appeared on the library science at a time when libraries were about to change from closed access. This coincidence of training was a major reason for its popularity and is a factor in its continuing success, having been translated into our thirty languages.

In the introduction of DDC 20th edition, it is given that DDC is used in 95 percent of public/ school libraries in the USA. Not only DDC extensively used for physical location of

material, it has also been used in many bibliographies, subject headings lists, indexing and abstracting services. It has undoubtedly reinforced the popularity of the Scheme.

1) Revision and Administration of DDC: In the span of 120 years (that is since the publication of its first edition in 1876), the DDC has been continuously revised to keep pace with the knowledge. A new edition has been released every six years approximately. The actual responsibility for the development and revision of DDC lies with the Decimal Classification Division at the Library of Congress, whose work involves the classification of works for whole MARC records and LC catalogue cards. The newest revision of DDC21 is the 'Electronic Dewey' published in 1996, which is the result of scholarly research, careful analysis of current literature, and consultations with users.

2) **The New Edition DDC21:** The latest DDC21 has many interesting features. The OCLC Forest Press, Ohio (USA), offers the scheme in two convenient formats: 1) Print Edition - A traditional four volume set; 2) A New Microsoft Windows Version - Dewey for Windows.

3) **Important Changes in DDC21:** There are three major changes in the classes 350-354 Public Administration, 370 Education and 560-590 Life Sciences. In Tables, some changes are made in the Standard Subdivisions. Christianity have been relocated in order to reduce the Christian bias. These adjustments reflect political and social changes, such as the major revision of Areas Table 2-47, for the countries of the former Soviet Union. New subject areas that have gained momentum of library warrant. Ever since its 20th edition, items such as 'Rap Music', 'Virtual Reality', and 'Snow Boarding' have been incorporated in the new edition.

4) **Added Features:** The entries on new topics, selected built numbers and terms to provide entry vocabulary for foreign users, have been added in the Relative Index. More inter-disciplinary numbers are included in the main schedules and Relative Index as well. The terminology has been revised and updated, increasing the scope of currency, sensitivity and global usage.

5) **Compact Disc:** DDC21 is available in Microsoft Windows based version on compact disc. Dewey for Windows lets you point and clock your way through familiar DDC functions and more with

   i) An easy-to-use windows format: drag and drop information between windows;

   ii) Expanded search and display options;

   iii) Ability to share a single CD-ROM among multiple users on a LAN

   iv) Provision to display multiple DDC records from the schedules, table, index and manual on one screen.

6) **Dewey for Windows - Special Features:** Over 4,000 new entries appear in the electronic index

   - the annotation feature lets you add notes to the schedules to reflect local classification decisions;

- sample bibliographic records show how DDC numbers have been used;
- the database includes benefit numbers from the Dewey Relative Index, and
- Cut and paste Dewey numbers into OCLC records using PRISM and passport forwindows.

The Dewey for Windows database includes separate records for built numbers that are included in Dewey Relative Index. Records for built numbers include index terms and the nearest matching DDC schedule number with its caption. Although Dewey for Windows cannot effectively provide complete numbers for all topics, more built numbers will be included in future electronic version of the classification.

The latest edition, 21st edition of DDC includes revisions, additions and corrections to Edition 20. 20th edition has the DDC updates. Data processing and Computer science have been extensively revised and the numbers expanded. Throughout DDC 20 common terminology has been used to replace often obscure and technical terms. Many minor but important revisions have been made in several subjects.

Due to criticism of Western bias, many changes and modifications have been made. Charles Marten of LC rightly remarked that DDC is a "system bound up in and made to lift notation, and the notation to fit the classification system".

DDC21 still remains the ultimate tool for organising the library collections and it is the world's most widely used classification system. The electronic Dewey where the scheme's structure allows the classifier to trace hierarchical relationships. Numbers are linked to a maximum of five LCSH terms and the searcher can enter a key word and main class and be directed to the class marks, related class marks and the corresponding LCSH terms. The approach is in key words, phrases or class marks and online help. The DDC's plans would seem to ensure the scheme at the leadings edge in the interaction between classification and technology maximisation of the effectiveness of the Electronic Dewey.

4.2.2 Library of Congress Classification (LCC)

The Library of Congress was set up in the year 1800 to provide a reference collection for the Government of US in Washington, D.C. As the collection grew to almost a million books with an annual accessions of nearly 1,00,000 items a serious thought was given to develop a separate system of classification suitable to LC collection. Thus, developed a new system called, Library of Congress Classification (LCC). This scheme resembles that of Cutter's Expansive Classification (EC). In LCC the traditional disciplines are chosen as the main classes. Therefore, it has a book oriented basis, rather than philosophical. It also lacks scientific approach of theory of classification.

Most of the schemes appeared between 1899 and 1930. LCC was created by individual teams working on each class and within these teams, individuals would work independently on their own particular subject specialisations, under the coordinating control of a subject editor. This is a feature which has continued to the present day and accounts for one of the

unique aspects of LCC. Each main class is published separately and is virtually independent of the others. These factors have had an impact on LCC's subsequent development and its value as a general classification scheme. The scheme is arranged first into main classes and their sub-classes, where appropriate, and schedules for these are produced in groupings.

The LCC System

Main Classes

A	Generalia, Polygraphy
B	Philosophy and Religion
C - F	History
G	Geography
H - L	Social Sciences
J	Politics
K	Law
L	Education
M	Music
N	Fine Arts
O	Language and Literature
P	Science
R	Medicine
S	Agriculture
T	Technology
U	Military Science
V	Naval Science
Z	Bibliography, Library Science

Each of the classes is further sub-divided into its main divisions. For example, in class L Education, we have the following divisions:

L	Education
LA	History of Education
LB	Theory and Practice of Education
LC	Special Aspects of Education
LD-LG	Individual Institutions
LD	United States
LE	America (Non-US)
LF	Europe
LG	Asia, Africa, Oceania
LH	College and School Magazines and Papers
LJ	Student Fraternities and Societies. US
LT	Textbooks

Notation is further expanded by the addition of numbers used arithematically. Each of the divisions has available to it the notation 1 - 9999 for further enumeration of subjects and these are read arithmetically with gaps left for expansion. Eg:

HV 6254 Offences against the Government (General

HV 6373 National General Works

HV 6275 Treason

The alphabets I,O,W,X and Y are spared for the future accommodation of new areas. Decimal extensions are of the recent use to accommodate new subjects. The use of LC class number is further extended by Cutter's Author Number and sometimes by the year of publication leading to lengthy numbers.

LC class marks are not expressive of subject content and make no attempt to display the hierarchy of subjects. Therefore, they do not provide a symbolic language where the notation can reflect the subject content but rather provide simply a locationary device. The result is that the notation conveys no sense of the structure of the scheme and the relationship and hierarchy of subjects is often difficult to discern from the schedules. Equally the lack of detailed summaries or breakdowns of classes in the schedules makes it difficult to find one's way around the subjects without resources to the Indexes. There is no single index to the LCC. Each separately published part of the schedules has its own index and these can in themselves be fairly extensive.

There have been attempts at the cumulation of these separate indexes, one produced by the Canadian Library Association, the other compiled by Nancy B. Olson, consisting of multiple volumes (15). A useful publication which appeared in 1992 was the first edition of a Manual to guide classifiers in the use of the LCC scheme.

One of the great strengths of LC lies in the fact that classificationist and classifier are the same. The scheme is, therefore, a superior example of a pragmatically developed system of proven success and value to the many libraries using the scheme. It provides minute details in specification and a more scholarly approach in many disciplines than the more populist DDC. Although enumerative, it serves the needs of the academic sector more satisfactorily than any other popular general scheme at the moment. It is estimated that 60 per cent of the US libraries are using this scheme. Its class numbers appear on MARC record and CIP data.

Revision

The process of automating LCC was encountered with a number of problems, mainly relating to lapses in the scheme, which have already been explained to you above. The scheme is constantly revised, and the changes made after a great deal of deliberations are announced in its quarterly publication, *"LC Classification - Additions and Changes"*. The weekly edition of LC, *Information Bulletin* also carries the changes.

4.2.3 Universal Decimal Classification (UDC)

Universal Decimal Classification (UDC) was designed as a practical general classification scheme, Paul Otlet and Henry La Fontaine in 1885, the founders of International Institute of Bibliography (IIB). IIB is presently known as International Federation for Information and Documentation (FID). UDC is chiefly a bibliographical schemes, that is designed for the indexing and description of the contents of document rather than for the physical arrangement of a

collection. It is, therefore, based on and aimed at the organisation and retrieval of information from all kinds of literature and in particular the provision of the details necessary to handle pamphlets, reports and periodical literature, wherein analysis could be carried to an almost extreme fineness. These features have remained an integral development of UDC and has greatest impact in the classification of special libraries, documentation centres and information bureaux, where it may be found in use for shelf arrangement and in the arrangement of the classified catalogue. A range of abstracting and indexing services, particularly in the field of science and technology, either arrange entries in their classified sequence in UDC order or carry UDC numbers on the entries.

UDC was initially conceived as an expansion of the basic structure of DC, which would explicit the internationally comprehensible qualities of DC's notation. Permission for this was granted on the condition that the first 100 class and divisions would remain identical. However, from that original common ground the two schemes have developed in very different ways and considerable differences now exist between the main divisions of the two schemes even at a basic three or four digital level. UDC developed considerable synthetic qualities.

In the introduction to the International Medium Edition (IME) of 1985, UDC acknowledges itself to be a 'hybrid' of the enumerative approach in the Main Tables, where the primary notation for subjects is listed, and of the analytico-synthetic elements which are available via the use of the Auxiliary Tables. The UDC has evolved from a project to develop an enumerative into a faceted scheme of classification. It is also considered as the first faced classification, which helped Dr.S.R.Ranganathan to develop the method further. It is available in

1) Full Edition (4th ed. in English) lists 2,10,000 concepts

2) Medium Edition 1985-88 lists 70,000 concepts, while the

3) Abridged Edition, 1961 (3rd Ed in English) lists about 20,000 concepts. The scheme published by the British Standards Institution as BS:1000 is available in small fascicules too.

The UDC is one of the first schemes to introduce the concept of facet analysis and auxiliary tables.

Revision and Administration of UDC

Since its first publication in 1905, UDC has been owned and administered by the FID who controlled the revision process via a fairly complicated and slow, but continuous, careful and well-vetted, consultative committee structure. Individual and widely desperate national bodies were responsible for the actual publication of schedules (the English Standards Institution produced the English language schedules) and although international development was thus necessarily uneven, the scheme is available in various manifestation in 22 languages. Editions can be full, medium (30 %) or abridged (10%) and schedule publication. The Task Force set up in 1986 for UDC system development has made several recommendations as to creation of a 'Standard Version' in machine-readable format at the medium level. A new Code of Practice for revision was also drawn up to standardise the process.

English Medium Edition

This Medium Edition, published by FID is available in English, French and German languages. The English Text, BS:1000M (London:BSI), 1985-88 is in two parts: Part-1: Alphabetical Subject Index to Part-1. The Part-1 contains the Schedules of nearly 70,000 concepts in the order of basic classes 0/9, including Preface, Introduction and Auxiliary Tables. It is also available in the electronic version form. The Online database does not really equate with the full schedules, but comprises about 60,000 entries. The electronic UDC is accompanied by two guides, a manual explaining how to use the file for the classification of materials, and which will expand upon the introduction to the existing International Medium Edition, and a guide to the ways in which the file can be put to use.

The notable changes in the new medium edition are of vacating Language Class 4 to merge it with Class 8 Languages, Linguistics and Literature. The Class 4 is now spared for new subjects, falling between basic classes of 3 Social Sciences and 5 Natural Sciences. The 38 Commerce has been merged with 33 Economics. The entire subject, Sociology has been classed at 301. The Classes 34 Law and 5/6 Science and Technology are dealt in detail and expanded. For the subject Space Science an extensive schedule has been developed and incorporated at 629.7. Two new signs, double colon (::) and square brackets [] have been added to represent 'additions' and 'relations'.

Current Evaluation of the UDC

As has already been noted, UDC is the most synthetic of general scheme of classification. Although used internationally, UDC in various forms is available in 22 languages — there still exists in the scheme a Western bias, inherited. This is particularly evident in the class 'Religion', which is dominated by 'Christianity' and also present in the treatment of political ideologies and cultural variations. The Scheme is used all over the world, the peak of its popularity is found in certain European countries and in Russia, where it has been used extensively. Equally UDC has been the most popular classification scheme in special libraries and information centres through out the world.

A related advantage of UDC lies in its independence. It is, therefore, possible to use UDC class marks as a complement to post-coordinate indexing language techniques, to represent subject concepts, to act as a thesaurofacet or to display hierarchical structures as the automation held great promises for UDC in the sense that the UDC is the most suitable of the general schemes for retrieval of information in electronic files. Thus, the UDC may be a way to make more suitable for the online age without drastic structural revisions to the existing UDC. At the moment UDC is at a dynamic and fluid point in its history, where organisational changes having already been implemented. The future for UDC is likely to include, attempts to render the scheme more attractive and effective for online retrieval of information, the provision of a sounder classificatory base, the identification and targeting more specially of present users, and the development of strategies to encourage new users to adopt the scheme.

4.2.4 Colon Classification (CC)

The Colon Classification, popularly known as CC was first published in 1933 by the Madras Library Association, authored by restless, insistent and uncompromising Indian National Research Professor of Library Science and also called the *Father of Indian Library Movement*, Dr.S.R.Ranganathan (1982-1972).

The Colon Classification adopts a main class structure, but thereafter, within each main class operates on fully faceted lines. Each class has a varied number of facets and the order in which these are combined or cited is controlled by means of a facet formula based upon Ranganathan's view of the Fundamental Categories.

The Scheme is little used, but it has, however, been immensely influential, both on the development of theory and on the making of new special classifications. The ideas and the language of CC are also percolating into the other general schemes, to a greater extent. The CC is an analytico-synthetic scheme of classification which enumerates broad conventional subject areas. CC abounds with ideas and features.

CC has been changed quite rapidly from edition to edition because of the insistence on accuracy of subject specification and on keeping pace with the introduction of new compound or complex subjects and also because of the sheer intellectual fertility, inventiveness an investigatory zeal of its creator. The introduction of new and radical theories has been a common occurrence in CC, such as the development of PMEST as a citation order. Citation order is controlled by means of a facet formula. Ranganathan argued that all elements relate to one or the another of the five fundamental concepts: *Personality, Matter, Energy, Space and Time*. In citation the categories present are represented in this order, often written as PMEST. Each of the facets is introduced by punctuation marks, serving as facet indicators, which identify the notation and the facet to which it belongs. Thus, in CC a comma heralds Personality. Matter is prefixed by semicolon and Energy by a colon (As you are aware that originally the only punctuation used to connect the facets and hence the name of the scheme). Space is introduced by a fullstop.

Revision and Administration of CC

The first edition of CC was designed as the trial and error effort, as it was not based on any underlying principles in which the schedules and facets were built. Hence, he tried to evolve a theory of classification through CC scheme. The result was the publication of Prolegomena to Library Classification in 1937 based on this theory. The second edition of CC was published in 1939 with illustrative examples. From this Ranganathan started productive research on the subject for improving the scheme. The third edition appeared in 1950 without major changes. A facet formula for each basic class was provided. The term facet replaced the phrase 'train of characteristics'. The fourth edition has radical changes over its previous editions. Five indicator digits to indicate five fundamental categories were introduced. A number of Greek letters for partially comprehensive subjects were introduced. The fifth edition of CC appeared in the year 1957. In this edition several changes and things were added. This edition allocated zones for different kinds of isolates. Further, Ranganathan described the postulational approach to classification and used () parenthesis to represent subject device. Many Greek letters were introduced to expand the base of main classes. Further, second level of space and time isolates were also introduced. The sixth edition was published in 1960 with the following modifications:

1) Greek letters which were introduced in Edition 5 were avoided, except 'Δ ' Delta.

2) Indicator digit inverted comma was used to represent Time facet.

3) The concept of empty and emptying digits was evolved.

In 1963, reprint of sixth edition was published with the following corrections and amendments:

i) 'X' was employed as emptying digit. Therefore, the main classes HZ, KZ and LZ have been replaced by HX, KX and LX.

ii) Evolved the methodology for designing a depth schedule on the basis of refined techniques and principles.

iii) Evolved the concepts of four zones with 40 sectors.

Finally, the long awaited seventh edition of CC (CC7) with substantial changes from the earlier edition appeared in 1987.

Seventh Edition

The seventh edition of CC is in five parts, namely,

A Introduction
B Guidelines to the Beginner
C General Rules
D General Divisions and Common Isolates
E Special Isolates

One significant feature of this edition is the provision of Environmental Divisions and Common Property Isolates which are not found in CC6. The concept of fundamental category, Matter has been changed. It may manifest itself among the facets of a compound subject as Round-2 and so on. Let it matter-Method (M-Mt), or Matter-Property (M-P), Matter-Material (M-M). Consequent to this change, the facet formula for example for Medicine now is: Medicine-Organ-Property-Action, that is, L;[P];[MP]:[E]. The common auxiliaries such as (ACI), (PCI) and common facets Space and Time have been greatly enlarged. There is no general index to the schedules.

Current Evaluation of the Colon Classification

Some see CC as difficult to grasp. This is a problem which is more apparent than real. It provides not only a remarkable degree of precision in specification but also control of the required order, all delivered in a unified and cohesive package. The notation offers interpolation to a marvelous degree. By listing essential elements from which compounds can be made, the schedules are kept slim. The CC scheme would seem in a sense to be in competition with UDC rather than LC or DC in that its approach and application to order on shelves raise issues of simplicity of notation. Ranganathan's belief that single scheme could serve various purposes is wellknown. The obvious legacy of CC is seen in modern faceted systems and in much classificatory research. Less obvious is the challenge posed to other systems.

On a global scale but many, CC is used by a few, greatly admired by some. Not everyone is convinced of the value of faceted classification in any context. In India the work of developing CC goes on, it is used in a variety of libraries and it is studied in detail at every level of library science education. Ranganathan really sought precision akin to computer analysis in a pre-

computing stage. The other general schemes are investigating their own suitability for assisting the process of automated information retrieval and yet this is an area where research with CC has not been carried out, despite the fact that many of its features would lend themselves to the exploitation of the electronic medium.

4.2.5 The Bibliographic Classification (BC)

Henry Evelyn Bliss (1870-1955) believed for many years that libraries needed a more erudite system to in more intellectual respectability in the eyes of subjects specialists and of educators. His long evolving system was eventually published between 1940 and 1953. The special features of BC include alternative location for certain themes where expert views might differ, short notation, and some selective linkings of pure and applied sciences. BC has been used mainly in Great Britain and the Commonwealth countries. It is now being revised as faced classification within the much praised original BC outer shell or structure; this process too has been long evolving, from 1969 todate.

H.E.Bliss laid the foundation of his work in two large tomes. These seek to establish the credentials for the classification which as to come and also recognise that reconstruction must be preceded by some demolition. The Organisation of Knowledge and the System of the Science (1929) concerns the structure of knowledge, showing his considerable debt to the ideas of the scientists and philosophers. The organisation of knowledge in libraries published in 1933 discusses the principles of bibliographic classification, notation and what he saw as the faults of existing systems. His thoughts can be traced back at least as far as 1910 and yet BC (or BC1 as we should now call it, as there is a considerably revised BC2) was not completed until 1953.

Revision and Administration of BC

In the lifetime of Bliss and soon after, there was some interest generally in the BC and it was adopted as a scholarly scheme by some libraries, some of these being new services able to make a fresh start with impunity. The majority of these users are British libraries using the full BC. The publishers, H.W. Wilson, provided a revision bulletin and in the late 1960s two enthusiasts, Scott and Freeman, produced an abridgement for use in school libraries (referred to as the ABC, Abridged Bibliographic Classification). It were to serve the library and information world of the future, however the longer term position and to be secured. Accordingly, a Bliss Classification Association was set up in 1967, and a particularly keen enthusiast, Jack Mills, initiated plans for second edition of the scheme, to known as BC2. This revision as based upon the retention of the much admired order of the main classes, but with the key concept of through-going facet analysis within each main class added to those features which had been considered essential in BC1.

Four or five of the desirable features of must be examined in BC. First, we have the concept of Consensus, the scientific and educational consensus as Bliss called it. This simply a way of acknowledging that an effective classification should be based upon the way in which subjects are taught in colleges and universities. Bliss thought that the classificationist could find and act upon this consensus within each subject field. He also contended that the consensus is relatively stable and tends to become more so as subjects and disciplines become traditionally fixed. BC1 produced a useful and scholarly order within many classes. This was helped by a second principle, that of collocation and subordination.

Collocation merely means the bringing together of groupings which have strong relationships into close proximity in the final order. Subordination means, of course, the careful placing of each specific theme in due subordinate position to the appropriate general subject. However, Bliss also used the term 'Subordination' in a more specialised sense, in his development of the idea of gradation by speciality. This concept reflects the influence on Bliss of the great French writer and classificationist, Auguste Comte, who had developed a system in the 19th century. The idea of gradation by speciality suggests that although a number of topics may be equal in rank, some are in a sense more specialised in that they draw on the findings of others, being therefore dependent upon these other subjects.

According to Bliss, no single order could ever satisfy everyone nd a third memorable feature of BC1 is the provision of alternative locations, where they were sanctioned by scholarly authority, which he built into the scheme. He thus provided two or more locations for certain subjects, one would be chosen and the others left blank. For instance, Economic History can be subordinated to General History, or can go instead under Economics. Likewise, in BC1, Bliss preferred Religion to be at Class P, where its association with History, Ethics and Social Work are stressed by contiguity, but he provided an alternative alongside Metaphysics in Class A, where Religion schedules could be developed as they are in Class P.

Bliss provided various schedules in BC1 for this, some were of general application, others for specialised use only. Some of these schedules were drafted on late in the development of BC1 and not unnaturally with their quest for classification in great detail which much detail is require, they frequently sit rather uneasily alongside the idea of brief notation. The basic notation gives short and distinctive classmarks, with some scope for literal mnemonics, although Bliss was careful not to disturb the chosen order to create these. The quest for helpful order within each main discipline has been revolutionised by rigorous facet analysis, using a standard citation order whenever possible. There are more alternative locations than before, systematic schedules for recurring themes, such as form of presentation of Geographical Area, have been improved and some recasting of Generaliá Class.

A basic feature, is the promotion of retroactive synthesis by means of an inverted schedule. This merely means that the major facet in any class comes last in that class, so that when qualified by earlier facets, the notation becomes a retroactive one, which will indicate the appearance of a new facet in the classmark by a reversion to an earlier letter of the alphabet.

Current Evaluation

On the credit side of this scheme, strenuous efforts have been made to incorporate the best original BC principles and provide, within the shell of the original, a fully faceted and predictable modern system. There is ample advice in the schedules and elsewhere on applying the scheme, the Bliss Classification Association in Britain remains active and Tony Curwen edits the BC Bulletin. On the debit side, users have been lost to the scheme over the years because of revision delays, as the work on this comprehensive revision began at the start of 1970 and could not be completed even after two decades of time. As it is, the pattern of a predominantly one-person scheme, taking far too long to complete looks all too sadly like a repetition of the problems encountered in BC1. BC2 represents as good as a place as any to study the role and potential of classification in general, as well as faceted classification in particular.

rinally, the future does hold the promise of interesting developments. It also indicates the high esteem in which BC is held and the many good qualities it displays, particularly its structure, rapid adherence to and observance of faced principles inherent logic and the clarity with which it is explained.

4.3 SPECIAL CLASSIFICATION SYSTEMS

Having considered the systems, which attempt to cater for all fields of knowledge, we shall now look at schemes which take a narrower view, focussing their attention in some way. The field of special classification system is more vital than that of general classification in that a greater number and variety of interesting projects have been and are being developed. A special scheme is a smaller entity, both easier to envisage and to carry through. They provide a scheme which is aimed at a particular user group, or they cater for a form of material which requires special treatment.

Special classifications are particularly useful and often deliberately created for use in special libraries, they clearly have no value in public or academic libraries, except where such libraries has special or specialist collection. The special systems are designed to suit the specialised areas such as Petroleum, Machine Tools, Medicine, etc. as these serve the purpose of compilation of bibliographies, brings out abstracting and indexing services, subject headings lists, etc.

4.3.1 Concept, Need and Scope of the Special Systems

The library classification theory experts have defined the special classification systems variously. Ranaganathan defined it as "Scheme designed for depth classification of micro subjects, going only with one and only one specific subject field". Librarian's Glossary defined it as "a scheme of book classification which is applied to a section of knowledge". Special classification systems are the classification schemes designed exclusively with depth schedules. Thus, a special classification scheme is a system with a micro focus. Basically, they attempt to (1) provide a classification for an intensive or narrower areas of subjects, a discipline of study or an area of professional practice; (2) provide a scheme which is aimed at a particular user group; or (3) they cater for a form of material which requires special treatment.

It is the particular emphasis on the special subject field which is unlikely to be reflected by the general scheme. What is most commonly needed in a special library is a detailed classification for the major subject of interest and a broader treatment for all of the rest of knowledge. For example, in a special library in the subject "Oil Industry", the main focus will obviously be on "Petroleum Engineering" and the related technologies, as represented in the collection. General classifications are often unwieldy in use in special libraries, because even if they do provide sufficient details, the resultant notation is likely to be excessively lengthy. The general scheme may also suffer because of its size as revision and updating is notoriously slow. The special library is going to suffer more in these circumstances than the general, as their collection will reflect more quickly the full range of need subject coverage. The needs of a special group of users can often best be met by a scheme's flexibility in not prescriptively determining citation order. This is a common feature of special schemes. Special systems also exist to deal with particular physical forms of material, which require an unusual arrangement or are likely to be approached in a different manner by users.

4.3.2 Some Examples of Special Schemes

There are a large number of special schemes in existence. Some are very limited in usage and others have become popular among a number of libraries. There are schemes which have been created by librarians or curators of a collection, devised by an organisation to provide access to or to arrange its documentation, and those which have been invented to organise bibliographies. Interested non-professional enthusiasts have also developed their own, just as Jefferson did in the scheme which was first adopted and developed by Library of Congress. Some of the select sample schemes which exist with brief descriptions of subjects covered are given below:

1) ***Social Services Libraries' Classification Scheme:*** This is a fairly straightforward treatment of a restricted subject area.

2) ***Moy's Classification Scheme for Law Books:*** It is designed by Elizabeth Moys for Butterworths, who published a very wide range of materials in the field of Law (including the LEXIS databases of legal information).

3) ***Iconclass - Iconographic classification system :*** It is a scheme designed for the accessing and retrieval of pictoral images rather than documents.

4) ***Hitlist - a Classification Scheme and Thesaurus for Housing Information:*** It is published by Sheffield City Council. The existence of this scheme reflects an awareness of a subject poorly treated which as a topic would likely be scattered amongst a number of disciplines, such as architecture, social services and so on.

5) ***CI SfB :*** It is a faceted scheme for use in the organisation of all kinds of documentation relating to construction projects.

6) ***Thesaurofacet:*** It is a combined faceted classification and thesaurus, designed by and produced by English Electric company.

7) ***Thesaurus of Play Terms:*** It is a thesaurus with an associated classification relating to all aspects of children's play. It is an interesting and unusual subject for classification.

8) ***Classification Scheme for Railway Company Archives :*** In this scheme has a special role for the information is clearly envisaged and again the special needs of documentation and archival material are indicated.

9) ***Scheme of Classification for Careers-related Information :*** It is published by the Institute of Careers Officers, an organisation which evidently felt that their subject area had special requirements. In professional practice subjects often display a quite different relation to each other from that which is found in their academic study.

10) ***Classification Scheme for Adult Education :*** published by the Institute of Educational Librarians.

11) ***Classification Scheme for Small Museums:*** produced by the Museums Association of Australia. It has a restricted subject coverage and also restricted to a particular size of collection.

The most effective way of discussing special classification is simply to examine some schemes in a little more detailed manner.

4.3.3 The London Classification of Business Studies (LCBS)

The London edition of this scheme by Vernon and Long was published by the London Graduate School of Business Studies in 1970, after it has been tested in the Scholl's Library for a number of years. In 1973 a programme of revision by a small working party was established to produce a second edition which would allow updating of the scheme and the expansion of certain classes. No further significant revision or new edition has appeared despite the expressed hope in the introduction to the second edition (LCBS2) which was published in 1979. Based upon the literary warrant, the scheme has also drawn extensively the expertise of a large number of practioners as well as some utilising sections of BC2. From this wealth of expertise it emerges a combining classificatory theory with pragmatic effectiveness.

In 1979, at the time of publication of the second edition, there were considerable users of the scheme, ranging from a graduate school of business administration, institutes of management, banks and industrial companies. The wide geographic spread of these libraries was particularly noteworthy and encouraging for the creators of the scheme.

LCBS was the second business studies classification to appear, its only predecessor, the Harvard Classification of Business Studies (1960), is an enumerative scheme, while LCBS was designed upon analytico-synthetic principles. It consists of a number of classes, the notation for the foci of which may be combined to specify compound subjects. The scheme itself does not dictate the facet citation order, leaving that decision to individual libraries and classifiers with a reminder to record decisions for future consistency. Clearly this flexibility is designed not to discourage users by too dogmatic and restrictive an approach, this means, however, that there is no assurance of standardisation in usage of the scheme. LCSB does, however, suggest a perfect order of citation, which may be adopted in order to avoid inconsistency.

The schedules for the classes provide an abundance of details and these have the added advantage of forming a thesaurus which can be used in complication of subject index in support of the shelf order. There are three main categories, each of which contains a number of classes and a set of auxiliary schedules.

Management Responsibility:

A	Management
AY	Administrative Management
AZ	The Enterprise
B	Marketing
BZ	Physical Distribution
C	Production
D	Research and Development
E	Finance and Accouting
F	Personnel
G	Industrial Relations

Environmental Studies

J	Economics
JZ	Transport
K	Industries
L	Behavioural Sciences
M.	Communication
N	Education
P	Law
Q	Political Science
R	Philosophy, Science and Technology

Analytical Techniques

S	Management Science
T	Operational Research
U	Statistics
V	Mathematics
W	Computer Science
X	O & M and Work Study
Y	Library & Information Science

Auxiliary Schedules

1	People and Occupations
2	Products and Services
3/4	Standard Subjects
5	Geographical Divisions
6	Time
7	Form

As can be seen, only certain subjects are included. The editors being very conscious of the need to avoid the temptation of becoming over-general, and other disciplines which in any general scheme would have their own place in the arrangement, such as mathematics, library & information science, are dealt as though a subdivision of business studies. This is of value to the majority of users, as this is how they view the material required, but one can see ho a distorted view of the universe might be promulgated in a user who is thus encouraged in his or her vision of mathematical science as a subsidiary aid to business management.

As we see from the above outline, the notation is comprehensible and conveys order reasonably well as it is chiefly alphabetic, although standard subdivision tables use numbers. It is not consistently expressive of hierarchy and facets are joined simply and in all instances by a slash '/' which although is fails to identify the nature of the facet to follow, does ensure simplicity of users. Examples:

(1) Trade Unions in the Oil Industry

Trade Unions	GC
Oil Industry	KTC
Classmark	GC/KTC

(2) Labour Relations in England in the Motor Industry

Industrial Relations	GA
Motor Industry	KHB
England	5111
Classmark	GA/KHB5111

(3) The Use of Sampling Techniques in Market Research

Market Research	BD
Sampling	UD
Classmark	BD/UD

The schedules are very well supported by an able and informative introduction by K.G.B. Bakewell, who revised the scheme for its second edition, and by the helpful and profile notes throughout. Overall this scheme has a very important contribution to make in the burgeoning field of business studies.

4.3.4 INSPEC Classification

INSPEC Classification is a classically simple, enumerative scheme, which provides enormous details in specification within four broad classes:

Section A	Physics
Section B	Electrical Engineering and Electronics
Section C	Computers and Control
Section D	Information Technology

These classes represent the published sections of the Science Abstracts and the schemes's main function is to provide for the physical arrangement of these abstracts in the classified sequence of the publication, rather than shelf arrangement. In the introduction to the published schedules, attention is also drawn to the use of the classification as an aid to subject retrieval online from an electronic database of records. The use of the scheme is largely seen, then, as an aid to subject retrieval whether from a printed source or from an electronic database. The user interested in neutron physics turns first to the subject index, where they find the class mark A2820 and then consulting the schedules, find attention directed to related subjects and notes indicating the extent of use of that classmark.

A2820 Neutron Physics
(see also A2540 Nucleon-induced Reactions and Scattering)
1977-1973-76 Use A4610; Use a1243

The final note this entry indicates what will be restricted by use of the notation, when the notation is combined with a data of publication as part of a search. This latter feature is necessitated by the fact that as INSPEC has grown it has proved necessary to allow for ever-grater specification of detail in order to cope with the older material has not been reclassified as it would in most library collections, understandably enough, given the many millions of records available via INSPEC online. Such a reclassification would be mammoth task and

could not be applied retrospectively to a printed index, the value of which is precisely as a retrospective resource. In information retrieval too, especially in such a rapidly changing field, searchers are often interested in material of specific period, most commonly recent material, and therefore it is helpful to combine the classmark with a time-span.

The chief advantages of INSPEC scheme are its provision of enormous detail and special subject concepts such as the mass ranges peculiar to Physics and the quasi-hierarchical structure of its notation. The latter allows for the effective use of truncation in an electronic environment, when, for example, one can identify all subdivisions in the hierarchy.

4.3.5 Moys' Classification Scheme for Law Books

It. is a refreshing to come upon a classification so unabashed about its focus upon books as to name these in its title. The first edition of this scheme appeared in 1968, with a second edition in1982. It is the work of a single author Elizabeth Moys who has also produced a Manual of Law Librarianship and, as such, it is an example of the excellent work which can be accomplished by a dedicated and knowledgeable individual. The field of law is a complex subject area with many very special characteristics. Such as the significance of jurisdiction, in its classification. Documents often have to be classified as a piece of legislation, interpretation, discussion or case study. The concept of division into primary and secondary materials must, therefore, be acknowledged in the classification of law. The scheme has, as a consequence of its creator's special understanding of the literature and its users, proven to be very popular.

One might have expected that law, as an ancient and revered subject of study, would by now have developed a classification system reflecting its significance, but it is rather sad to note that, in her Introduction to the second edition, Elizabeth Moys declares that consultations were mainly with law teachers and law librarians.

Moys' system has several interesting features, the most unusual of which is the provision of two class marks for each subject listed in the schedules. This has been done largely in acknowledgement of the fact that often general libraries want and need the extra detail and features of a special scheme and Moys, therefore, utilises a notation which will allow libraries using the two most popular schemes: DC and LC, to fit the scheme into their existing arrangement at the appropriate class, 340 for DC and K for LC, without affecting the use of all the other classes. Even in the library may, therefore, choose to employ the rest of LC or DC for other subjects as treated, thus allocating them a more representative place in the universe of knowledge than would be achieved by their subordination as an aspect of Law. However, Moys does offer the use of a subdivision KZ for other subjects for special libraries who prefer a more economical approach and do not wish to purchase the complete editions of LC or DC.

For each subject as found in tne schedules, then there are two classmarks provided, for example:

Subject	LC	DC
Ancient and Medieval Law/		
Byzantine Law	KE	343

General works	251	.5
Collections	255	.51
Basilica	256	.52

so that for the subject of general works on Byzantine Law, there are two options: KE251 or 343.5

The approach is fundamentally enumerative with respective to jurisdiction, with all the common law jurisdictions grouped together with the individual countries, such as Scotland or the US or the Channel Islands, being specified by the application of notation from a special table. The basic order of classes is therefore, General and Non-national Legal Systems (including International, religious and ancient)

Modern national legal systems

Common Law
Treatises
Other modern legal systems

Non-legal subjects

Beyond this basic enumeration, synthesis is utilised fairly extensively via such tables as the one described above and others which specify elements, for example, the legislative nature of materials and the subjects of law.

At all times, there is indicated in this scheme an awareness and understanding of the use to which the classification will be put. Since the second edition, revision has continued and amendments and additions are published in the 'Journal of Law Librarianship'.

4.4 PRINCIPLES AND DESIGN OF SPECIAL SCHEMES OF CLASSIFICATION

A special classification scheme must start from a very clear perspective of the role it is going to play within the library. The first question we must ask ourselves is 'why ?',

Why is the scheme being created ?

Will it be used for shelf arrangement or for subject retrieval?

Why are existing schemes not catering sufficiently for the particular demands of the subject ?

Do users have particular needs which are not being met ?

Does more detail is especially desirable ?

A full understanding of the needs of the likely users of the scheme is essential, of their most typical approaches to information and of their attitudes to their discipline. Equally the scheme should be based very firmly on literary warrant, that is, a firm knowledge of the kinds of document with which the classification will have to cope. Many special schemes of excellence have thus been developed for use with a comprehensive collection, such as the FIAF classification for film and television which is published by the International Federation of Film Archives, or they may have been created for the arrangement of an extensive bibliographic tool such as the INSPEC Classification created for the subject arrangement of bibliographic record in Science Abstracts. In both instances, these build very effectively upon such an awareness of the literature of the subject.

4.4.1 Principles of Special Classification Schemes

A special clasification scheme needs to be based upon sound classificatory principles, although as Gilchrist has noted successful schemes have been devised without conscious resource to classification theory. While special classification schemes may be either enumerative or faceted in nature, the latter approach is the more common. A significant number of excellent special schemes are based upon synthetic principles, such as the Construction Industry's CI SfB - an early faceted scheme with a sound and continuing development programme, and in ways, it is in the field of special classification that such principles have been most enthusiastically stated upon.

4.4.2 Designing a Special Scheme of Classification

What if there is no scheme which caters for our special needs though we decide to go it alone ? We might choose to embark upon the creation of our own special scheme. There are certain considerations which must be borne in mind before we rush into the not inconsiderable task of inventing a classification scheme, even if it is only a special scheme. The initial work of developing schedules may not in itself be burdensome and, given that no appropriate scheme exists, that such work may according to principles best suited to the collection and the library must be alert to the need to continue to update and revise the scheme. One danger lies in the relegation of such a task to an individual in the library service, who may not bear in mind the need to establish and record the principles and the method of practice in both using and continuing to develop and expand the scheme. The establishment of policy, codes of practice and a manual of use of the scheme, for both classificationists and classifiers, are therefore essential.

Aids to the creation of a special scheme are available and anyone attempting such a task would, of course, be well advised to familiarise themselves with the principles of classification. There are certain stages in the construction of a special classification scheme, which are detailed below:

Stage - 1 : Identification of Concepts

The first stage in the process is to identify the subject terms used in the field to be covered. For this purpose it is best to work from the literature and from the literature of the kind with which the scheme will be used, that is, if only books are to be arranged then the terms should be chosen from books, if journal articles then these should be examined; if reports literature are to be included then they should form part of the base from which the system will

be developed. The intention here is to build up a select base of terms, from which the subject area can be analysed into its component facets, and that base should be built upon literary warrant.

A significant sample of pieces of literatures should therefore be analysed into its discrete subject concepts. This sample should, as noted, reflect the kinds of documents which will be held in the collection and should also represent not just the most recent publications, but should reflect the subject as it has developed over a period of years, thus avoiding a slated view of the significance of current issues and concerns. where one has an existing developed collection to work from, then there is no difficulty in identifying materials where a new collection is being developed then it may be necessary to work from subject bibliographies or indexing services.

Stage - 2 : Analysing the Concepts into Facets

The concepts must be analysed or arranged into the category or facet to which they most naturally belong. Some facets will be quite obvious and readily conceived, others may cause some mental turmoil.

The facets have thus far been identified from the list gathered together. Facets in any classification which wishes to provide full specification of detail and, given that the random list from which are working are all books, it also illustrates the complexities of treatment which are to be found even at that level of publication. Most schemes would not seek to cater for this level of detail. Each of these facets must then be fully extended, identifying and listing existing and potential foci as comprehensively as possible.

Stage - 3 : Arrangement of foci within Facets

The equal foci or individual concepts, must be arranged in the most helpful order within each of the facets.. Sometimes a logical order will present itself very obviously, at others there will be no evident reason for adopting one order over another. Another principle which might be applied could be that of developmental or revolutionary sequence. Or we may consider using what is termed canonical order, that is, an order which is traditional and generally acceptable to users of the scheme. Also, the Wall-Picture Principle which may be applied suggests that the relationship between subjects should be considered in determining order. This principle is in many instances chronological, but may not always be so, if one were dealing with disease and preventive medicine, then one might argue that the prevention comes first, the disease may not appear at all, but disease must have existed before prevention could be considered necessary.

It is not helpful to stick rigidly to a chronological sequence when that would separate subjects which are related in other ways. Rather than create entirely new facets, where possible it is perfectly reasonable and more economic to base the commoner facets such as space and time upon those of existing schemes such as BC or UDC.

Stage - 4: Establishing the Facet Citation Order

The question to be considered when establishing the first facet in our citation order is which of the facets is the most important in terms of ensuring that works dealing with any given focus within the facet should be collected. The degree of collocation of works will decrease

in facets placed lower in the facet formula or citation order. In order to establish this, it is necessary to examine each of the facets, establishing the priority of the significance of each. We should use as our guide the literature of the subject and also the preferences of the typical user of such a collection. How is the subject studied? Ho do people in the industry approach material? How do interested readers look for material? Clearly, the answers may vary depending on the organization which the library or collection will serve. One can of course avoid the issue by saying that the choice of facet formula is left to the classifiers, the library itself. The advantage of course is that it is possible to ensure that the optimum order achieved for the particular environment in which the scheme will be used in practice.

Stage - 5 : Creating the Schedules

The next step in the creation of a classification scheme is the recording or writing out of the schedules. The simplest method of applying the principle of inversion, at this stage, is to record the facets in reverse order, creating an inverted schedule. Then the notation can be simply applied in a logical filing sequence and the desired shelf order will be obtained when the facets are combined. Here decisions have to be made about what the notation will hope to achieve, whether it is to be expressive or simple, how facets will be introduced and so on.

The final stage, the importance of which should never be underestimated, is that of providing an index, catering for all of the relationships not revealed by the chosen citation order to the schedules, together with instructions, either in an introduction or manual, as well as where necessary in the schedules, which will aid the classifier in the use of the scheme.

Stage - 6 : Testing and Evaluating the Scheme

It is all very well to develop a scheme according to the best principles, but before full implementation it is vital that the scheme be tested, initially in order to identify any problems. This can be done with a fairly small number of records in order to ensure that the principles have been applied correctly and to see what the final shelf order looks like. After all, this must continue to be evaluated for effectiveness so that revisions and improvements can be carried out as necessary. Inevitably, as the collection and the literature grow, there will be expansions and developments which were not foreseen and it will be necessary to expand upon classes, your scheme must therefore be monitored constantly in the light of these new appearances.

4.5 FUTURE USE OF CLASSIFICATION SCHEMES

Much effort is being made and there is a definite commitment to providing automated versions of major schedules but, without wishing to detract from the quality of research and the progress made to date, much more is required to determine the benefit of the approaches currently being pursued. There exists in practice a fundamental dichotomy between the use of classification as a shelf ordering device and its use as a retrieval tool. The many relocations and complete revisions of sections of any of the major schemes which would be required to enhance the scheme's online retrieval potential are simply not practical. Notational complexities which would be required explicitly to denote superordination and subordination of topics to assist online browsing would almost certainly undermine the effectiveness of schemes as shelf location devices.

Librarians must be able to demonstrate not only that information retrieval can be enhanced using classification but also that it is worth making a significant monetary investment to take it work. Included in this cost would be the requirement to reclassify stock to ensure that the most recent version of the classification scheme being used is uniformly implemented throughout the library and the necessity to ensure that the reclassification is conducted in a timely manner in order to keep pace with any revisions of the published classification schedules. Currently many librarians would argue that it is financially impossible to maintain large bibliographic collections to keep pace with schedule revisions.

The purpose of an information storage and retrieval system, if one looks beyond a basic functional description, is essentially to answer questions. This presupposes a dialogue between the system and the user of the system and the structure of the classification scheme should be able to form a framework for entering into this dialogue. One of the most interesting results of recent experiments using classmark searching online to provide additional records is that it in fact provides different records from those provided using keyword approaches. The problem of how to provide access to these records without increasing the number of false drops incurred to an unacceptably high level, however, remains resolved.

Results of research currently being undertaken confirm that central issues which must be tackled are the inadequacy of indexing languages, the problems associated with use of Boolean logic and the need to support browsing. More theoretical research in the field of information retrieval generally concludes that changes in the technology we use to access information systems have outpaced work on the principles used to store and access information.

There is no perfect indexing language or means of describing documents which will produce adequate content description of documents. The attempt to index by subject is in practice impossible as we cannot find a lacuna is. The emphasis on constructing information retrieval systems must therefore be moved to describe the relationship between documents. Each new document added to the collection must be treated not as a discrete entity but must be processed in a manner which explicitly links it to other related documents in the information system, allowing the document to be present not only in a single hierarchical structure but in an interconnected web of information nodes. It would appear, therefore, that classification which by definition seeks to expose relationships between documents, far from being redundant in online.

4.6 LET US SUM UP

A general classification embraces all knowledge and enables us to classify macrodocuments according to the subject content in them. Examples are DDC, UDC, LC, CC, BC, etc. A special classification restricts itself to a part of knowledge and enables us to the extent of minute classification of knowledge in classes belonging to a conventional subjects such as physics, chemistry, machine tools engineering, etc as well as interdisciplinary subjects. It helps to organise the micro-documents such as monographs, research reports and articles in periodicals.

Most of the general schemes of classification came into existence at a time when the concept of documentation and information services were unknown. Their use was restricted mainly to the arrangement of books on the shelves. The advances in the universe of knowledge on the one hand, and the developments in the documentation techniques on the other, extended the application of classification. The general schemes, mainly enumerative in structure, lacked the necessary resilience and flexibility to take up the additional responsibility. Their defects and inadequacies became too glaring which applied to the purposes of classification, especially microdocuments. The special schemes of classification allows for minute classification of subjects in the specialist field chosen. It attempts to provide a helpful arrangement of subjects as required by the specialists in the field. It also represents the concepts and terms used and sought by the specialists in the field.

4.7 REFERENCES AND RECOMMENDED BOOKS

BRITISH Standards Institution. *Universal Decimal Classification* (BS1000M:1985). London: BSI, 1985.

CHAN, Lois Mai. *Immroth's guide to the Library of Congress Classification,* 4th ed.. Englewood: Libraries Unlimited, 1990.

CLASSIFICATION for information retrievall edited by KGB Bakewell. London: Bingley, 1968.

CLASSIFICATION of library sciencel edited by Inder Mohan Goswami. New Delhi: Commonwealth Publishers, 1995.

FOSKET, A.C. *The subject approach to information,* 4th ed. London: Bingley, 1982.

KHANNA, J.K. and R.Vohra. *Handbook of library classification systems.* New Delhi: Beacon, 1996.

KRISHAN Kumar. *Theory of classification.* New Delhi: Vikas, 1979.

"NEW deal for UDC". *FID News Bulletin,* 41(11);1991.

ROWLEY, Jennifer E. *Organising knowledge: An introduction to information retrieval.* London: Gower, 1987.

SATIJA, M.P. "Use of colon classification". *International Classification* 13(2); 1986. pp.88-92.

WIESTHUIS, G.A. and S.Bliedung. "Thesaurification of UDC: a preliminary report". IN *The UDC: essays for a new decadel* edited by Alan Gilchrist and David Strachan. London: Aslib, 1990.

4.8 ASSIGNMENTS

1) Identify a special area of knowledge where a special scheme of classification is not available and design a scheme of classification suitable to it.

2) Give a brief account of the problems faced by general and special classification schemes with possible solutions through new designs, methodology or principles.

4.9 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) Trace the recent developments in the organisation of classes in UDC.

2) Critically examine the merits and demerits of the seventh edition of colon classification

3) Discuss the need and purpose of special classification schemes and briefly describe any one special classification scheme you have studied.

II SHORT NOTES

a) Library of Congress Classification

b) Bibliographic Classification

c) Electronic Dewey

UNIT - 5 : UNIVERSAL DECIMAL CLASSIFICATION (UDC)

Structure

5.0 AIMS AND OBJECTIVES

This unit introduces to you the Universal Decimal Classification, one of the most widely used classification scheme through out the world.

After going through this unit, you will be in a position to

- trace the genesis, development, features and principles of UDC
- describe the structure and features of notation of UDC
- discuss the variety of auxiliary tables and signs employed in UDC
- explain the wider usage of UDC and the manuals brought out to promote the use of UDC in LICs.

5.1 INTRODUCTION

UDC is one of the most important and general schemes of library classification. The scheme was developed on the basis of Dewey Decimal Classification and first published in the year 1905 entitled *Classification Decimale Universelle* by Paul Otlet and Henri La Fontaine. This scheme is revised and updated from time to time by International Federation for Information and Documentation (FID). The scheme is very widely used in most of the Special Libraries and Information Centres and also by bibliographical services through out the world. The scheme possesses certain distinct features and notational techniques so as to meet the ever growing needs of the user libraries and information centres. Let us, therefore, discuss the qualities of notation, auxiliary tables in UDC and its wider use for classifying and systematic arrangement of documents on the shelves and entries in bibliographies, and Indexing and Abstracting Bulletins.

5.2 GENESIS AND DEVELOPMENT

The credit of using decimal fraction notation for the first time in library classification goes to Melvil Dewey, the author of Decimal Classification published in 1876. The value of this scheme was soon recognised and its use rapidly spread to the continent of Europe. In the year 1895 the first International de Bibliographie (IIB) was founded. During the course of its history, the Institute changed its name twice. In 1931 it was changed to Institute International de Documentation (IID). Again in 1937 the Institute assumed the present name of Federation International de Documentation (FID). The FID General Assembly in Montreal, 1986 agreed to make change by incorporating the word 'information'. Since then it is known as International Federation for Information and Documentation although the acronym FID continues to be used.

The IIB in the beginning sponsored a scheme which was initiated by two Belgian enthusiasts Paul Otlet and Henri La Fontaine for the establishment of a comprehensive classified Index to all published information to which people all over the world would contribute and which would in turn be available to all. After examining the existing systems Otlet and La Fontaine concluded that Dewey Decimal Classification which was then in 5th Edition offered the most suitable basis, being a purely subject classification using internationally known Indo-Arabic numbers as decimal fraction notation which offers infinite hospitality in the subdivision of a class.

Otlet and La Fontaine wrote to Dewey and sought his permission to extend the details of his scheme to make it suitable for arranging the kind of Index they had in mind. Dewey gave permission to extend the schedules on the condition that the order of the main classes and divisions be maintained and that maximum computability in development be obtained. These two Belgians then proceeded to enlarge the schedules of DC by adding extensively to its enumerative classes. The extension provides an apparatus for synthesis or number building. In 1905 the complete International edition in French of what was later called the *Classification Decimale Universelle* was published. This edition consisted, of some 33,000 subdivisions with an alphabetical Index of 38,000 entries titled as *Manuel du Repertoire Bibliographique Universal*. This provided basic classification schedules for the great card subject Index compiled in Brussels by the two Belgians.

The First World War (1914-1918) was a heavy blow for the activities of the Institute and it was not until the mid-twenties that work on the second edition was resumed. Otlet and La Fontaine supervised the Humanities and Social Sciences and F.Donker Duyvis supervised the Natural Sciences. From 1927 onwards, full International Editions embodying fruits of international cooperation in additions and revisions appeared in French, German and English together with numerous abridged editions in these and other languages. F. Donker Duyvis became the Secretary General of IIB in 1929 and continued in that post until his retirement in 1959.

UDC as we know today owes a great deal to this great man Donker Duyvis, for his tireless efforts to make it one of the great general bibliographical classification schemes. UDC found another enthusiast in S.C. Broadford, Librarian of the Science Museum Library, London. In his book on Documentation, Bradford exhaustively deals with UDC which he applied in his Library. The first work on UDC was published in English by Bradford, an abridged schedule used in the Science Museum Library entitled *Classification for Works on Pure and Applied Science in the Science Museum Library*, the third edition of which appeared in 1936. In course of time British Standards Institution (BSI) became the official British Editorial Body for bringing out UDC in English.

5.2.1 Different Editions of UDC

As pointed out earlier, the first complete edition was published in French in 1905 as *Manual du Reportoire Bibliographique Universale*. The second complete edition, also in French, was published over the years 1927-1933 had the title *Classification Decimale Universelle* and indication of the change of emphasis since the first edition. It contained some 70,000 classes enumerated in its tables and substantial alterations were made in the main classes Science and Technology. By this time it became the most detailed and flexible classification scheme ever published for use in special libraries of every kind. In 1934 the third full edition in German, entitled *Dezimal Klassification*, was begun and was completed in 1952 making it volume index was published in 1951-1953. The BSI which took the responsibility for the publication of UDC in English began in 1943 and the publication of 4th edition and was expected to be completed by the end of 1977. It constitutes a British Standard BS 1000. Other full editions under preparation include revisions of the German and French editions and new ventures in Japanese (1950), Spanish (1955), Polish (1959) and Portuguese (1961).

In addition to the above mentioned full editions, abridged editions of the whole UDC have been published in 13 different languages. The first British abridged edition was published in 1948, the second edition in 1957 and third edition in 1961. Multilingual edition BS 1000B in German, English and French was published in 1958. These are three separate Indexes. A supplement to this has been published covering the years 1958 to 1968.

Another development in the publication of UDC was the bringing out of medium editions and special subject editions. Medium edition as first published in German, was intended to fall between the full and abridged editions. This medium edition contains about 30% of the full tables. The International medium editions in German and English versions were expected to appear in 1977 with the French editions following shortly after. But the International Medium Editions (IME) English Text was brought out by BSI in 1985 (BS 1000 M: 1985) in two parts, viz., Part 1: Systematic Tables (1985) and Part 2 : Alphabetical Subject Index (1988). (Please refer to Unit which deals with salient features of IME). Special subject editions were brought out in Nuclear Science, Mining and Metallurgy and Building. These editions are usually based on the practice of a larger library system.

5.2.2 Organisation and Revision

The development and maintenance of UDC is the responsibility of FID. This is exercised through its international committee on which all national member committees are entitled to be represented. The overall control is vested in the Central Classification Committee FID/ CCC consisting of FID Secretary General and editors of full editions of UDC. Each member nation of FID may have its own national committee for bringing out editions in the language of that country. In addition to these committees there are national and international subject committees. The International Committee report to the FID/CCC and the national subject committees report to the national committees.

Proposal for revision or extension generally comes from the users of UDC. A proposal forwarded through the national committee or by individual users is first considered by the editors who are in constant touch with FID office at the Hague where an upto date minute master copy of the complete UDC is maintained. The proposals are then sent to the headquarters where they are studied carefully to see that it does not clash with any existing or proposed schedules. If the proposal is accepted it is published as a *P-note*. These proposed alternations are made public and lie on the table for four months during which time any user of UDC may comment on them. If no comments are received during this four month period, the proposal is deemed to be accepted and the schedule becomes part of UDC.

Every year P-notes are progressively cumulated into three year volumes. The process of revision in UDC is continuous and it is both a source of strength and weakness. The new schedules proposed by the user is scrutinized by FID/CCC to ensure that the proposals are sound from the classification point of view. The procedure that is followed, before accepting any proposed schedule as official, is a long drawn affair. It takes nearly two to ten years for accepting a proposal and in the dynamic and fast growing disciplines like Sciences and Technologies the proposed schedules tend to become out-of-date before they are accepted as official and incorporated in the master schedule of UDC.

5.2.3 Management of UDC

The UDC managed by the organisation founded by Otlet and La Fontaine known since 1937 as the Federation International Documentation (FID). The official languages of FID are English, French and German. FID works in close co-operation with various national organisations which in turn have various consultative arrangements with users of the scheme. The classification division of FID maintains a complete master version of UDC which consists of the text of the French edition of 1927-33.

The published editions of UDC in various languages are issued by the member organisations in various countries with the authority of FID. The Scheme is under continuous revision. The proposed amendments known as P-notes are circulated by FID to subscribers for comments and suggestions. The suggestions and amendments received from the users and experts are considered by the FID. The accepted suggestions, amendments will become the authorised versions of FID when published. The published amendments are incorporated into the language editions of UDC. The published amendments become known as *'Extensions and corrections'*.

The English editions both full and abridged of UDC is being brought by British Standard Institution (BSI) from time to time. BSI operates a committee structure in which interested organisations are represented for formulation of revision proposals.

5.3 FEATURES AND PRINCIPLES OF UDC

5.3.1 Features

It must be stated that UDC is a practical scheme based on the demands of pamphlets, reports and periodical literature rather than on the framework of a theory. Although the scheme is based on DDC, it can claim to be the first 'analytico-synthetic' library classification. It lays more stress to achieves co-extensive class number, i.e., detailed specification than the achievement of a sequence of subjects for optimum helpfulness. It avoids a lacunae of numerous classification schemes by providing a standard system covering all the disciplines and may be used in any type of general or special library. It is a general classification scheme and it is not a bundle of special classification schemes but an integrated whole.

The Scheme reflects exhaustive enumeration in the schedules with due provision for synthesis or coordination. It is amenable to adjustments to meet special needs. This is possible because the citation order in any given class allows alternative treatments. The use of synthetic devices like Colon (:) permits coordination of concepts in different permutations, thereby minimizing the rigidity in the enumerative classification schemes. The notation in UDC is international in nature so that a file organised by it makes sense in any language. An international body for its maintenance and revision with the full cooperation of its users guarantees the continued existence of the system as a current and up-to-date one. The terminology used in UDC helps as a comprehensive vocabulary of terms with indexing purpose.

5.3.2 Principles

UDC is based on the following principles:

1) It is a classification in the strict sense depending on the analysis of idea content, so that related concepts and groups of concepts are brought together and the arbitrary and often haphazard systematization of alphabetical and other arrangement is avoided.

2) It is a universal classification in that an attempt is made to include in it, every field of knowledge, not as a patch work of isolated, self-sufficient specialist groupings, but as an integrated pattern of correlated subjects.

3) It is a universal decimal classification constructed on the principle of proceeding from general to the more particular by (arbitrary) division of the whole of human knowledge into ten branches, each further sub-divided decimally to the required degree.

4) It is essentially a practical system for retrieval of information in which the order of subjects is not of much importance than the provision for detailed specification.

5) It also accepts the principle of mutually exclusive classes, collocation of related subjects and consistency of approach.

6) To a certain extent it has tried to remove national and racial bias by removing these factors and preferring common facets. But still it is an essentially Western-oriented system. It neglects oriental religious systems, philosophies, cultures and social systems.

7) Its notation consists of Indo-Arabic Numerals used decimally which allows infinite hospitality to subordinate classes.

8) It employs certain notational techniques by which it is possible to link simple main class either with other main number or with auxiliaries indicating place, time and similar commonly used categories.

5.4 STRUCTURE OF UDC

As this scheme is based on DC the order of main classes and main divisions is almost the same that of DC. The notation is slightly different and there is no three figure minimum in UDC. Thus 5 is Science, 53 Physics and 531 Mechanics. The outline of the main classes denoted by decimal fractions is as follows:

0 Generalities: Methodology, Documentation, Scripts, etc.,

.1 Philosophy, Metaphysics, Logic, Ethics, Psychology

.2 Religion, Theology.

.3 Social Sciences: including Statistics: Law, Education.

.5 Pure Science, Mathematical and Natural.

.6 Applied Science; Medicine and Technology.

.7 The Arts, including Architecture, Photography, Entertainment and Sports.

.8 Literature, Belles-Letters.

.9 Geography, Biography, History.

After the publication of International Medium Edition (IME) in 1985, the main class Linguistics has been transferred to Literature main class 8. In the outline of main classes the final zero is not used to give three figure minimum. In certain cases, it is used with some meaning. For example:

3 Social Sciences

33 Political economy, Economics

330 General concepts of Economics

5.5 NOTATION OF UDC

The notation consists of Indo-Arabic numerals used decimally. In addition, several other signs are used to introduce common facets. The classes are divided from general to specific. The following example will reflect this process of division.

.3 Social

.34 Jurisprudence, Law, Legislation

.343 Criminal Law

.343.8 Penology

.343.81 Penitentiary establishments, institutions

.343.819 Other penitentiary establishments

.343.819.2 For women and young girls

For the sake of convenience, the initial decimal point is omitted. A point is inserted at every three digits in the class number purely as a visual aid. When .00 and .0 auxiliaries are used in the class number or for some mnemonic purpose such as parallel division of 91 and the (4) to (9) auxiliaries, e.g., 91.44 Geography of France, the position of the point varies.

5.5.1 Qualities of Notation

The following qualities are evident in the notation of UDC. These are briefly described below:

1) **Subordination to Order:** The notation reflects the process of division from general to specific and subordination to order. The above example demonstrates this order. But because of enumerative nature and limited base, the class number for minute subjects is lengthy.

2) **Simplicity:** The main schedules are derived on the basis of Indo-Arabic numerals alone. But for common auxiliaries and special auxiliaries a number of indicators were used

which make the notation mixed. Thus the notation undoubtedly becomes complicated in nature. As a result of the use of several indicators, a definite ordinal value has to be awarded for each of these non-numeral symbols.

3) **Brevity :** As stated elsewhere, the notation is far from brief. This quality suffers due to small base of ten Indo-Arabic numerals as decimals and poor allocation of numbers to classes. Many class numbers in UDC exceed six digits in length. Because of poor allocation, extensive and dynamic subjects got less number of places than relatively restricted and static subjects. The purpose of several synthetic devices is to achieve co-extensive class number, and as a consequence the class numbers tend to be very lengthy.

4) **Hospitality :** A notation of the scheme should be hospitable to new subjects. There are two types of hospitality, viz., hospitality in array (coordinate classes) and hospitality in chain (subordinate classes). Hospitality in array is achieved in UDC by using 'Centesimal' device: Alphabetical device is also used, e.g., names of plants, persons, places, industrial products etc., to extend any array. Sometimes zero is also used to represent array of classes. e.g.: 308 Sociography; 408 kinds of languages and dialects. Infinite hospitality in chain is achieved due to decimal fraction notation. This is also possible due to several synthetic devices with different signs which give the notation faceted quality.

5) **Flexibility:** Another notable quality of notation in UDC is its flexibility in the sense that it allows alternative arrangement of subjects. This quality is achieved:

i) by the use of distinct signs as facet indicators so that the facets may be in different orders

ii) the use of colon (:) as a general linking sign

iii) the use of intercalating devices - for general intercalation, ( ) and " " for intercalating space and time facets

iv) the point of view numbers. The following example shows the hierarchies and flexibility in constructing class numbers using different citation orders.

63	Agriculture
632	Plant diseases, pests
632.9	Pest control
632.93	Methods
632.934	Chemical Control

Example: *Chemical control of pests in India*

632.934 (540) Pest Control - Methods - Chemical Control - India
or 632.93 (540)4 Pest Control - Methods - Chemical Control - India

632.934 (540) Diseases, Pests - Control - Methods - Chemical Control - India or
632.9(540)4 Diseases, Pests - Control - Methods - India - Chemical Control or
632.9(540)34 Diseases, Pests - India-Methods - Chemical Control or
632 (540)934 Diseases, Pests - India - Methods - Chemical Control or
63(540)2.934 Agriculture-India-Diseases, Pests-Control Methods- Chemical Control

5.5.2 Synthesis and Mnemonics

The scheme provides the most powerful synthetic apparatus available in series of common and special auxiliaries which act as systematic mnemonics. These auxiliary tables are supported by a series of signs of combinations and indicators.

The following synthetic devices are employed in UDC:

Symbol	Name	Significance
+	Plus sign	Documents dealing with two subjects
/	Stroke	Documents covering several topics which are consecutive in UDC
:	Colon	Relationship between two subjects
[ ]	Square Brackets	Relationship sign used when one of two subjects is obviously of secondary importance

Common facets:

=	Equal sign	Language subdivisions
(0...)	Parenthesis	Subdivisions of form
" · "	Inverted Commas	Time facet
(1/9)	Parenthesis	Space facet
.00	Point double Zero	Points of view

Special Auxiliaries:

-	Hyphen
.0	Point Zero
'	Apostrophe

In the strict sense of the word all the above do not represent synthesis. The use of alphabetical device does not at all represent synthetic principle. But the above signs and auxiliary schedules do provide extra-notation strength for achieving co-extensive class number.

5.5.3 Starvation Policy

The notational techniques in UDC have greatly contributed to the flexibility of scheme. Its bias towards Science and Technology is clearly evident from the schedules in AEE and IME. In order to meet the requirements specially in Sciences and Technologies starvation policy has been introduced for future development of the scheme.

Starvation policy means that a piece of notation which has been superseded by a new schedule is to be left unused for ten years. After ten years the said notation can be used again with completely new meaning. Changes in the overall structure of scheme can be possible if there are vacant notation in which to be accommodated newly emerging subjects. This is the reason for transferring Class 4 Linguistics to Class 8 Literature in IME. After ten years, the vacant notation 4 can be used for new disciplines.

5.5.4 Bias in the Schedules

You are aware that UDC is based on the 5th Edition of UDC. Therefore it has inherited the basic structure with all its faults and prejudices. Consideration efforts has been made by the promoters of UDC to neutralise this Western bias. But they could achieve very little success. The best example of bias in UDC is Religion main class. Christianity has been given undue importance neglecting other major and oldest religions. UDC which is intended for world wide use cannot continue with such bias, which is totally unacceptable. In the IME you will find that the digits 22/28 have been allocated for various aspects of Christianity and only 29 has been allocated for other religions of the world. The same is case with the main class Philosophy. Oriental philosophies especially branches of Indian Philosophy find no place in the schedules of AEE and IME.

5.6 AUXILIARY SCHEDULES IN UDC

The auxiliary schedules provide maximum synthesis in UDC even though the bulk of the schedules are derived on the basis of enumeration. These may be divided into two groups, viz., the common auxiliaries and special auxiliaries. These are in fact, a set of common facets and facet indicators which enable to 'synthesise freely than the more rigid notation in DC. The following schematic representation will give a clear view of these two types of auxiliary schedules.

Main Schedules	*Auxiliary Schedules*
In which are set out;	In which are set out;
Main divisions 0 to 9	connecting signs and
and the various special	the common auxiliaries
-.0' (Table (k))	(a) to (i)

Some of these auxiliaries are restricted in application while others may be used frequently.

5.7 WIDER USE OF UDC

The use of UDC is quite impressive. Its present position in Europe is remarkably encouraging. Latest figures indicate that of a world total of some 110,000 libraries known to be using this classification, about fifty thousand of them are in USSR and another forty thousand in Poland. A scheme which has originated in Western Europe is now widely used in Eastern Europe. It is also widely used in Latin America and Japan. In France, Italy, Spain and Portugal, UDC is covering more ground especially in Technical and University Libraries. According to G.A. Lloyd, it has been estimated that some 20,000 copies of the Abridge Edition (BS 1000A) have been sold in countries as well as in English speaking countries. In the developing countries of Asia and Africa interest in the use of UDC is steadily growing. In India many special libraries use UDC for shelf arrangement and also in Indexing and Abstracting Services.

Another important fact is that the medium edition of the UDC published in German prepared by the German Classification Committee was intended to serve as a master edition for translation into other languages. It has been further estimated that about 170,000 copies of the full edition of the schedules in English language are in the hands of various libraries and its continuing influence is considerable. The efforts made by such experts' as J.E. Wright, Chairman, British Standard Institution's Committee OC/20/4 in keeping the schedules upto date by revision is also contributing to the widespread use of UDC by the libraries in the world. In addition to Libraries and Information Centres using UDC, several Indexing and Abstracting Journals and Services use UDC. J.Mills has studied the Abstracting and Indexing Journals and services using UDC.

5.8 SALIENT FEATURES OF ABRIDGED ENGLISH EDITION (AEE) (1961)

In Unit 4, you have been briefly introduced to the genesis, development and salient features and principles of UDC. This Section deals with the salient features of AEE. 1961. The AEE, 1961 (3rd Edn. Revd.) was bought out by British Standards Institution (BSI) in response to widespread demand for comprehensive short editions in English. This edition serves as a manual for classifying books and other documents in small and medium sized libraries.

The following sections deal with features AEE, nature of auxiliary notation, use of auxiliary tables and special auxiliaries. While constructing the class number you have to follow the citation order and filing order which helps you to arrange the books on the shelves. The Alphabetical Index appended at the end helps you to locate a class number in the schedules.

.1) Development of Abridged Editions in English

In response to the wide spread demand from several quarters for a comprehensive short editions in English, abridged editions are being brought by BSI, the official agency. The abridged edition BS 1000 A was first published in 1948. The second abridged edition with radical revision was brought out in 1957. The third abridged edition was brought out in 1961,

which took into account of the amendments listed in 'Extensions and Corrections to the UDC' during the four years 1956-59.

.2) Contents of AEE

The AEE, 1961 provides a brief note, on DDC and UDC. It also provides general introduction to UDC followed by history and availability, development and revision, principles, structure, notation, auxiliary notation, notes on using UDC and use of auxiliaries and compound numbers. After introduction you will find tables of auxiliaries, followed by outline of the main divisions. You will find detailed schedule for each of the ten main classes. At the end an alphabetical index is appended.

.3) Auxiliary Notation

In order to classify compound and complex subjects which require the addition of elements, attributes, aspects, point of view, AEE provides for a variety of auxiliary notation of special signs and numbers listed in one set of tables (a) to (k). These are applicable as needed to all parts of the schedules. These serve to distinguish not only the finer details and complexities of the subject matter or content of a document, but also the language or other form in which it is presented.

.4) Tables of Auxiliaries

Auxiliaries are means of eliminating repetition in all parts of the printed main schedules, in that the group contain recurring and general subordinate concepts such as language, form, place, time and point of view, etc. AEE provides for these tables of auxiliaries starting from page 10. The use of these auxiliary tables are briefly presented to you in the following subsections.

.1 Connecting Signs
.2 Common Auxiliaries of Language
.3 Common Auxiliaries of Form
.4 Common Auxiliaries of Place
.5 Common Auxiliaries of Race and Nationality
.6 Common Auxiliaries of Time
.7 Alphabetical and (non-decimal) Numerical Subdivisions
.8 Common Auxiliaries of point-of-view

.5) Connecting Signs

These appear in the scheme in the following order;

(a) Addition and Consecutive Extension sign = and/

The plus sign + (plus or 'and') may be used to connect the notation for two subjects which are commonly associated or the same broad subject in different aspects for which no single number exists, e.g., 32+35 Political Science and Public Administration; 575 + 613 Genetics and Personal Hygiene; and (540) + (510) + (520) India, China and Japan.

The stroke / meaning 'form ... to ...' is used to join consecutive UDC numbers denoting a range of concepts which collectively form a broad subject or branch of knowledge for which no single piece of notation exists. In the process of division, some important links have been missed. Then the / can be very useful to connect missing links, e.g.,

42/49 Languages and Dialects

69/68 Manufactures

(b) Relation Signs

These are : (colon), [ ] (Square brackets), and L:: (double colon)

Colon:

The colon sign is the most important of the connecting symbols and widely used to link two or more UDC numbers denoting related concepts of equal value. This colon and square bracket sign devices are provided on the model of 'divide like the whole classification found in Decimal Classification. This device is used to achieve a phase relation. For example,

633.1:551.58 *Effect of climate on Cereal Production*

331. 892:676 or 676:331.892 *Strikes in Paper Industry*

32:35 *Comparative study of Political Science and Public Administration*

The device may be used to combine foci from different facets of the same main class. For example,

025.5:027.021 *Information Services in Technical Libraries.*

361.92:362.8 *Flood Relief Work by Youth Welfare Clubs.*

631.86:633.61 *Organic Fertilisers and Sugarcane Yield.*

or this device may be used to enumerate the foci within a faced by using the schedule from another class. For example,

338.962:661.1(540) *Large Scale Glass Industry in India*

331.1: 669.14 *Industrial Relations in Steel Industry*

664.8:634.3 *Preservation of Citrus Fruits*

Square Brackets [ ]

This replaces colon when the second class number represents a subordinate concept of which 'no separate entry is required by reversal'. This is used as a means of 'intercalating', i.e., changing the facet order when the normal means of subdivision would be by means of the colon. For example,

015(540)[31] *Indian National Bibliography, Statistics of*

It also allows intercalation of any facet at whatever point in the citation orders required.

33[622.343]1.895 *Economics - Copper Mining Industry-Labour-Lockouts*

or 331[622.343]..895 *Economics-Labour-Copper Mining Industry-Lockouts*

620.172[669.13] *Material testing-Tensile test* (for cast iron)

or 620.1[669.13]72 *Material Testing* (for cast iron) *Tensile test*

In the 3rd abridged English edition of UDC it has been suggested that this device 'is refinement which may usually be dispensed with especially in printed or other widely circulated publications for which the: (colon) should be preferred'.

A.C. Fosket suggested that 'if join the UDC numbers by means of a + or: and then followed by this, by, say, a form division, it may be difficult to arrive at unambiguous subject statement. By using square brackets the ambiguity can be eliminated. For example,

09+294.3(05) *Manuscripts and Periodicals about Buddhism.*

[09+294.3] (05) *Periodicals about Manuscripts and Buddhism.*

[297:294](540) *Islam in relation to Hinduism in India.*

Double Colon (::)

Fosket suggested the use of double colon :: in certain situation while indexing a document. He states that 'Colon is widely used as a 'Pivoting' device for generating additional entries by reversing or cycling. Once the index has generated the original class number, cycling is a purely mechanical task which can be done by a clerical worker or a machine. There will, however, be occasions, when we do not want to reverse because the second part of class number is very clearly a subsidiary. The use of double colon is now suggested as a means of indicating this situation. For example,

635.9::582.734 *Ornamental Roses*

If the Indexer did not think it necessary to make entries in the Botany section for individual plants specified in Horticulture, double colon can be used as shown in the above example. This device eliminates to make additional entry under 582.734.

Common Auxiliaries of Language:

This is employed to specify the language in which the document is written taking language number from Class 4 and replacing the 4 by equals. This device is used rather as part of the description of a particular book than as part of the subject.

5+6(03)=20 *Encyclopaedia of Science and Technology in English*

22.05=9483 *The Bible in Telugu Translation*

633.88(038)=20=9143 *English-Hindi Dictionary of Medical Plants.*

Common Auxiliaries of Form (0...)

The form divisions correspond to the Standard subdivisions of DC but are used in Parenthesis. These are listed in some detail. It mostly consists of outer forms of presentation. Theory and Philosophy, study and teaching are excluded and (091) historical presentation is the only inner form. For example,

027.5(548.4)(094.5) *A.P. Public Libraries Act*

54(09) *History of Chemistry*

3(048) *Social Science Abstracts*

Common Auxiliaries of Place (1/9)

These divisions are based on the numbers from DC's 900 class. The schedule not only contains usual political divisions but also several other sub-facets of space such as zones, orientation, physical features, etc. These may be used as primary facets. Relationship between countries may be shown by the use of colon within brackets. For example,

027.7(540) *University Libraries in India*

327(540:510) *Foreign Relations between India China.*

When the geographical or regional aspect is emphasised entry may be under the auxiliary which the precedes the main number.

(540) 388.9 *Indian Air Transport Systems*

(520) 347.772 *Japanese Trade Mark Laws*

In certain cases the place can also be used as intercalating device, i.e., to be inserted into the middle of an existing piece of notation to change the facet order.

329.14 *Communist Party*

329(540)14 *Communist Party of India*

Orientation:

622.33(540-11) *Coal Mining in Eastern Regions of India*

634.0(540-18) *Forest Products of North Eastern Regions in India*

Political and Administrative Units

027.4(5484-202)	*Public Library Service in Rural Areas of Andhra Pradesh*
353(548.1-201)	*Municipal Towns in Tamil Nadu*

Physiographic Designations:

631.67 (282.5:548.4)	*Canal Irrigation in Andhra Pradesh*
633.17(952:548.2)	*Cultivation of Millet in the Semi-Arid Regions of Karnataka*

Common Auxiliaries of Race and Nationality:

These are based on the common auxiliaries of language and may be developed from the linguistic schedule. These serve to indicate racial aspects of a subject denoted by any preceding number. This facet is of limited application in nature. For example,

301.16(=914.3)	*Study of Social Relations among Hindus*
392.5 (=924)	*Marriage Customs of Jews*
325.48 (=96)	*The Rise of Independence Movement among the Africans*

Common Auxiliaries of Time "..."

These are used to indicate the period covered by a work or less frequency the date at which it appeared. Years, months and days can all be shown in logical sequence. Centuries are indicated by the use of two or three figures. Dates of B.C. have minus signs prefixed. There are also many other sub-divisions of time as seasons, months, days, hours and even minutes. The complete flexibility of time facet in UDC is very useful. A minimum of our figures is observed for single years, three for decades, two for centuries. For example,

"1981.01.26"	26th January 1981
"0727"	(A.D.727)
"0004"	(A.D.4)
"198",	the 1980's
"085"	the 85's,
"19",	the 20th Century
"01",	the 2nd Century AD.

As stated already, dates before the birth of Christ are prefaced by minus sign.

"-0032", 32 B.C., "-02", the 2nd Century B.C. For other period embracing several centuries, decades or years are denoted by the initial and final figures separated by the (/) sign. For example,

"04/14" Middle Ages

"18/19" 19th to 20th Century

Other aspects of time facet are accommodated in "3/7". "31" past, present and future. For example,

321.7(540)"313" *The future of Parliamentary Democracy in India.*

"32/37" include seasons, months of the year, peace time, war time, work time, etc. For example,

35.07(548.4) "372" *Government offices working hours in Andhra Pradesh*

The digit "4". Age measured in years. For example,
63.33 (548.4) "45-30/45" *The conditions of Agricultural labour 30-45 years old in A.P.*

The digit "5" represents time intervals and periodicity. The second digit shows the unit of measurement, e.g., "54" Monthly, "54-02" bi-monthly. 05(540)54"-02" Indian bi-monthly magazines. The digit "7" represents phenomenological division such as "77" spontaneous, e.g., 338.62:621.33 (548.4)"731" uninterrupted supply of electricity to large industries in Andhra Pradesh.

5.9 LET US SUM UP

The genesis and development of UDC is an important landmark in the history of classification schemes. Even though the scheme is primarily based on DDC (5th Edition), the promoters have included several notational techniques to achieve analysis and synthesis of a class number. The features and principles of the scheme have greatly contributed to attain international status.

The introduction of decimal fraction notation, several signs and symbols to indicate various aspects of a subject have greatly helped the synthetic nature of the scheme. The several common and special auxiliaries provided in the scheme greatly help to achieve coextensive class numbers.

The revision and updating policy of UDC and its publication in various abridged and full editions in important languages of the world have greatly contributed for its wider use and acceptability by libraries, and information centres through out the world.

5.10 REFERENCES AND FURTHER READING

BRITISH Standards Institution. *Universal Decimal Classification* (BS1000M:1985). London: BSI, 1985.

DHYANI, P. "Universal Decimal Classification International Medium Edition". *Library Review* 21; 1989. P.165-172

KRISHAN Kumar. *Theory off Library Classification.* 4th ed. New Delhi: Vikas, 1988.

5.11 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) Discuss the structure of UDC with suitable examples.

2) Compare the structure of UDC with that of DDC.

3) Write an essay on the restructuring of UDC and trace out the recent developments that are taking place in reorganisation of UDC.

II SHORT NOTES

a) Starvation Policy

b) AEE

UNIT - 6 : SUBJECT INDEXING

Structure

6.0 AIMS AND OBJECTIVES

Subject Indexing assumes an important role of an Information processing and retrieval system. This unit aims to provide an overview of subject indexing system.

After studying this unit, you should be in a position to

- explain the meaning of subject indexing
- describe various methods of subject indexing processes
- distinguish the generic and specific characteristics and explain their distinct functions, and also indicate the use made of them in different contexts.

6.1 INTRODUCTION

One of the important functions of an information retrieval system is to match the contents of documents with the users' information requirements or queries. This emphasises that the content of each input document in the collection is to be analysed and represented in such a way that it becomes convenient for matching. The process of constructing document surrogates by assigning identifiers to text items is known as Indexing and when the task of indexing is based on the conceptual analysis of the subject of documents, it is called *"Subject Indexing"*.

Subject Indexing is a method of information retrieval. Due to complexity in the nature of development of subjects and in the way information seekers go about it, access to relevant subject at any time needs efficiency in recall and relevance. Subject indexing in general leads to access through term control. The subject index should help leading a searcher from an unclear or a rough statement to an extensive standard one. The system of control of searching should be such that a searcher is helped to arrive at a statement compatible with the information base that is being searched. This is done by either broadening or narrowing down the concepts given in a statement of the searchers.

Indexing operations have been performed intellectually by human indexers for quite a long time. Automation systems have been developed comparatively recently where text analysis and indexing is performed by computers. However, the basic tasks involved in indexing is the same, that is to analyse the content of the given document, and representation of this analysis by some content identifiers or keywords. Lancaster mentions that the process of subject indexing involves two quite distinct intellectual steps, namely, Conceptual Analysis and Representation. Although the methods for representing the contents of documents vary from system to system, the analysis of the subject is the same in each case. In subject classification, the basic objective to arrange the documents on the shelves according to their subject contents. The result of the conceptual analysis is represented by some artificial language or notational symbols. A number of such systems, viz., DDC,UDC, LC, CC, etc. have been in use for quite a long time. (These schemes are dealt clearly in Unit-4: Classification Systems: General and Special)

In subject indexing, the basic objective is to match the contents of documents with the user's queries and thus, the product of the conceptual analysis of the subject is represented in natural language form. A number of systems, namely, Chain Indexing, POPSI, PRECIS, etc. have been developed over a period for preparing subject index entries of documents. In order to standardise the task of choosing appropriate keywords for generation of index entries, a number of vocabulary control devices have been developed. Examples of such devices include thesaurus, classaurus, thesaurofacet, etc. These tools help the indexer to choose the most appropriate term to represent the subject at the stage of indexing, and also help the users to pickup the most appropriate terms while formulating the query.

6.2 SUBJECT ANALYSIS

By the terms 'Subject Analysis', we mean the analysis of the thought contents embodied in a document. An author puts forward his/her ideas in a document, and it is task of the subject indexer to determine specifically what the author wanted to say in his/her work. Subject analysis means the presence, identification and expression of subject matter in document texts, databases, controlled and natural languages, information requests and search strategies. The most difficult part of subject indexing is in that phase where the indexer, who is not necessarily a subject expert, tries to summarise the contents of the whole document within a few words.

Certain guidelines have been put forward in the literature which guide an indexer in determining the content of a given document for the purpose of indexing. These guidelines are actually a set of questions which the indexer has to keep in mind while examining a given document. Once answers to these questions are found, the indexer gets an idea about the subject matter of the document.

Following are the questions set by BS:6529, which illustrate the general factors to be considered while determining the subject of a document.

1) Does the document deal with a specific product, condition or phenomenon ?

2) Does the subject contain an action concept, an operation or a process ?

3) Is the subject affected by the action identified ?

4) Does the document deal with the agent of this action ?

5) Does it refer to particular means for accomplishing the action, i.e., special instruments, techniques or methods?

6) Were these factors considered in the concept of a particular location or environment ?

7) Are any independent or dependent variables identified ?

8) Was the subject considered from a special viewpoint, not normally associated with that field of study, ie., a sociological study of religion ?

However, it may be noted that most of these steps require intellectual involvement of the indexer, and therefore, there is enough possibility that two different indexers may analyse

the content of a given document in two different ways resulting in two different index entries. In fact, this is a serious drawback of a manual indexing.

6.3 SUBJECT INDEXING

In an ideal environment of document retrieval, a document or a query statement is represented by a group of distinct index terms as well as the semantic relationships between these terms so that retrieval could be conducted on a structure of semantic relationships. Documents are retrieved on the basis of the correspondence between search terms expressed in a query index terms of the document. Subject indexing system, that is, indexing system based on the analysis of the contents of the documents, have been in practice in the retrieval world for quite a long time. Indexing systems, designed to assist in the retrieval of documents, operate by assigning index terms to the analysed subject of each document - either manually or automatically.

6.3.1 Pre-Coordinate and Post-Coordinate Indexing Systems

Subject indexing systems have been classified broadly as Pre-coordinate and Post-coordinate systems. It has already been mentioned that the major objectives of any indexing system is to represent the contents of documents through keywords or descriptors.

The following titles of documents may be carefully examined:

(1) Schools

(2) Colleges

(3) India

(4) History

(5) History of India

(6) History of Schools in India

(7) History of Schools and Colleges in India

The first four are single concept documents and their specific subject can be represented by a single term, i.e., Schools, Colleges, India and History respectively. The last three are two-concept, three-concept, and four-concept documents respectively. This means that the last three documents need two, three and four index terms or components to represent their subjects co-extensively. Due to inter-disciplinary nature of researches most of the documents, these days, cover composite subjects. The subject of each of these documents cannot be represented by a single index term. They require more than one index terms to represent them. For instance, in the seventh title given above, there are four index terms (i.e., History, Schools, Colleges and India), which are to be combined or coordinated to express its specific subject accurately. Unless the specific subject of a document can be represented by a single index term, as in the case of first four documents above, the combination or coordination of index terms becomes essential.

In fact, all the alphabetical indexing methods are coordinating index systems. They use concept coordination methods. They coordinate two or more concepts to represent a new one that differs from the concepts represented by the terms individually or in some other combination. If they do not do it, the content representation of the documents shall neither be specific nor accurate.

1) Pre-Coordinate Indexing

In the pre-coordinate systems, as the name implies, keywords chosen through the subject analysis stage are coordinated at the stage of indexing and thus, each entry represents the full content of the document concerned. Systems like, PRECIS, POPSI, Chain Procedure, Relational Index, NEPHIS, etc. are examples of pre-coordinate indexing systems. Entries prepared according to any of the pre-coordinate systems will represent the full context in which the entry words occur, whereas in post-coordinate systems entry under each term is generated without any context, ie., unless all the corresponding entries are found out, the content of the document cannot be known.

The index terms or components can be combined or coordinated in many ways. These will have to follow an appropriate sequence which may suit the approach of the majority of the users. Kaiser, Coats, Ranganathan, Farradane, Sharp, Austin, Neelameghan, Bhattacharya and others have contributed substantially towards evolving an appropriate sequence so that the index terms are coordinated at input level while building up the index file. The indexing systems they developed are known as Pre-Coordinate Indexing Systems, sharing the following common features:

(1) The subject-headings consist of two or many index terms

(2) The index terms are arranged in a pre-determined order, ie,the terms are pre-coordinated, and

(3) For each document profile, the index provides multiple entry points since a particular pattern of coordination may not meet the approaches of searchers.

Thus, Chain indexing, PRECIS and POPSI are pre-coordinate indexing systems because in them the coordination of index terms is done at the input stage in anticipation of user's approach. In all these, the subject of a document is analysed into constituent concepts and these concepts are then represented by symbols or words used in the indexing language. These symbols or words are finally arranged in a logical order following the syntax of the indexing language to derive the subject formulation. All these pre-coordinated subject formulations are maintained in an order in the index file as a key to the contents of the document collection. In sum, the main characteristics of a pre-coordinated indexing system are 1) they are based on subject analysis of documents, ii) they follow some specified citation order of terms, and iii) coordination of terms is done at the input stage.

Search Strategy

In the pre-coordinate indexing, the search does not involve problems. The user looks under the terms that he expects to find the subject described with and, with a good index, finds and follows instruction from his first entry point until appropriate documents references

have been retrieved. Both the indexer and the searcher should 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.

Limitations: Following are some of the limitations of the pre-coordinate indexing:

1) A particular sequence of coordination of index terms may not meet the approaches of all the users of index file.
2) To overcome the above limitation, multiple entry system and a network of references has to be provided. But even this covers only a fraction of the possible number of total permutations. Thus, the index file may fail to provide a particular combination which a user may be looking for.
3) This system requires both the indexer as well as the searcher to be thoroughly acquainted with the intricacies of the order of significance prescribed in the syntax of the language.
4) In view of the syntactical intricacies, this system requires intellectual involvement on the part of the indexer. This makes the processes time-consuming.
5) Pre-coordination of index terms may be too broad for a particular search
6) The user does not have any freedom to manipulate the coordination of terms in a desired manner to satisfy his specific needs.

2) Post-Coordinate Indexing

In post-coordinate systems, one entry is prepared for each keyword selected for representing the subject of a given document and all the entries are organised in to a file. When a user puts forward a query, it is analysed and some keywords are selected which are representative of the users' query. These query terms are then matched against the file of index terms and relevant documents are retrieved. Systems like Uniterms, Peek-a-boo, etc. are examples of post-coordinate systems.

The limitations of pre-coordinate indexing and consequent dissatisfaction with its working are responsible for the development of an alternative system which does not involve an order of significance. In this system, the concepts representing a composite subject are kept separately uncoordinated by the indexer. The coordination of component concepts or index terms is done by the searcher at the search or output stage. The information seeker has unrestricted freedom 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 index terms is done after the index file has been compiled, this indexing system is called Post-Coordinate Indexing.

The main characteristics of post-coordinate indexing are i) It is based on subject analysis, ii) no fixed citation order is maintained at the time of indexing, and iii) coordination of index terms is done by the user at the output or search stage.

It is clear that the process of concept coordination is there in both pre-coordinate and post-coordinate indexing systems. But in each the coordination is done at two different

stages. In the former, it is done at the input stage while preparing the index. In the latter, it is done at the output or retrieval stage. Bernier calls these two systems as 'Non-manipulative' and 'Manipulative' respectively. The post-coordinate index is designated as manipulative because it permits a greater degree of search manipulation and the index terms can be coordinated almost in any combination.

W.E.Batten (UK), G.Cordonnier (France), Calvin Mooers and Martimer Taube (both from US) have contributed extensively towards the development of post-coordinate indexing systems. Each one of them has devised a system, but Taube's system has been the most popular. Nevertheless, the principle on which all these systems are based is the same.

Term Entry and Item Entry Systems

The post-coordinate indexing system is of two types: i) Term Entry System, and ii) Item Entry System. The Term Entry System is that where term cards are posted with relevant document numbers. The Uniterm and Optical Coincidence methods are examples of this system. Item Entry System is that in which one card is maintained for a document. Each concept relevant to the subject matter of the document is represented in code form by a series of notches punched out from the holes along the margin of the card and the central area of the card is used for document description.

Exhaustivity and Specificity

The effectiveness of an indexing system is controlled by two parameters, called Indexing Exhaustivity and Term Specificity. By exhaustivity we mean the degree to which the subject matter of a given document has been reflected through their index entries. An exhaustive indexing system, thus, is supposed to represent the contents of the input documents fully. However, to attain this objective, the system has to select as many keywords as possible, to represent the idea put forward in the document. In a non-exhaustive system, only a few keywords are chosen which represent the subject grossly. Term specificity refers to how broad or how specific are the terms or keywords chosen under a given situation. The more specific the terms are, the better is the representation of the subject through the index entry.

Before discussing the impact of these two factors on the effectiveness of an information retrieval system, let us understand two more parameters, which are used to measure the effectiveness of an information retrieval system. They are Recall and Precision. Recall refers to the proportion of relevant documents retrieved by a system. Thus, Recall can be represented as

$$\text{Recall} = \frac{\text{No. of relevant documents retrieved}}{\text{No. of relevant documents in the collection}}$$

Precision refers to the proportion of retrieved documents which are relevant. Thus, Precision can be represented as

$$\text{Precision} = \frac{\text{No. of relevant documents retrieved}}{\text{Total No. of relevant documents retrieved}}$$

These parameters are expressed in percentages, that is both recall and precision may vary between 1 and 100 per cent. The objective of information retrieval system is to retrieve all the documents relevant to a query and simultaneously holding all those which are not relevant. Thus, a system will always attempt to attain both high recall and high precision.

Now, let us try to understand how these two parameters (i.e., Recall and Precision), are affected by the indexing exhaustivity and term specificity. Taking a simple example, say the query, 'Application of computational linguistics on indexing periodicals', if we choose any one term 'indexing', 'periodicals' or 'computational linguistics', we may lose a large number of relevant items. Now, if we take any to terms together, the system will retrieve more number of relevant documents. In other words, the more exhaustive are the indexing terms, the higher is the recall. Lancaster suggests that a high level of exhaustivity of indexing tends to ensure high recall, while Blair suggests that increasing the number of indexing terms increases the chances of hit, but the success is not guaranteed. However, by increasing the level of exhaustivity we tend to decrease the level of precision. This happens due to the fact that as we go on increasing the number of keywords, it may so happen that we choose such concepts which are very narrowly discussed in the given document. Therefore, although we shall be able to retrieve a large number of documents, the user may find very few among them with the subject matter at a desired level. In other words, the system will retrieve a large number of non-relevant documents. Thus, increase in indexing exhaustivity tends to increase recall but precision tends to fall.

6.3.2 Manual Indexing

The basic steps in the manual subject indexing process consists of -

1) Analysis of subject
2) Identification of keywords
3) Standardisation of indexing system
4) Choice of indexing system

- If the chosen system is a post-coordinate one then
- Preparation of entry under each term with reference to the document identification number;
- If the chosen system is a pre-coordinate one then,
- Preparation of an entry (main entry) using all the keywords organised in a way prescribed by the system;
- Preparation of index entries by using each significant term as an entry element and the full entry (main entry) as the significant terms in the main entry according to the rules prescribed by the system chosen

5) Filing of entries

It may be noted from the above that the first task in the process of indexing is to determine exactly what the given document is about, which is often termed as deciding the 'aboutness' of the document. The subject content of a document is sometimes referred to as 'intrinsic aboutness'. However, once the indexer collects information on the 'aboutness' of the document, his next task is to represent the same in a way suitable for matching with the users' queries. While doing so the indexer has to choose the appropriate keywords. We have already seen that the effectiveness of an indexing system and thus, the performance of the total retrieval system depends on the exhaustivity and specificity of the index terms.

A number of vocabulary control devices are available to guide the indexer in choosing the most appropriate index terms. Some pre-coordinate systems prescribe the use of vocabulary control devices which have been specifically designed for that system, e.g., PRECIS prescribes the use of an online/ inbuilt thesaurus, while POPSI prescribes the use of classaurus. The task involved in the post-coordinate systems are fairly simple, after choosing the keywords. The only task the indexer has to do is to draw index entry under each term separately along with the necessary reference entries ('see' and 'see also' entries). All the entries are then arranged in a file. The tasks involved in the pre-coordinate systems are, however, much more complex. The major problem relates to the coordination of index terms to produce meaningful entries. Each pre-coordinate system prescribes a number of guidelines which are to be followed for preparing the index entries. All these systems are based on some sort of categorisation principle, that is the index terms are first categorised, and then they are coordinated or arranged according to some rules prescribed for arranging the categories.

6.3.3 Automatic Indexing

When the assignment of the content identifiers is carried out with the aid of modern computing equipment the operation becomes automatic indexing. 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 original document texts, or document surrogates. The words occurring in each document are listed and certain statistical measurements are made, like word frequency calculation, total collection frequency, or frequency distribution across the documents of the collection. Before going into such details of the process, we should try to understand the advantages of automatic indexing. The following are the advantages of automatic indexing:

1) Level of consistency in indexing can be maintained.
2) Index entries can be produced at a lower cost in the long run
3) Indexing time can be reduced
4) Better retrieval effectiveness can be achieved.

PROCESS

Automatic analysis by means of word frequency analysis can be viewed as a two-tier problem. In the first stage, the problem relates to the identification of a technical vocabulary

characteristic of a given subject field. Once the vocabulary or index terms have been chosen, the second problem arises relating to the representation of the document with the help of the keywords.

However, coming back to the first problem, we may see that the idea of analysing the subject of a document through automatic counting of term occurrences was first put forward by H.P.Luhn in 1957. He proposed that -

* the frequency of word occurrence in an article furnishes a useful measure of word significance
* the relative position of a word within a sentence furnishes a useful measurement for determining the significance of sentences
* the significance factor of a sentence will be based on a combination of these two requirements.

The basic idea behind Luhn's theory was that more the frequency of occurrence of a term in a given document, the more significant is that term in denoting the subject content of the document. The steps for preparing indexes based on frequency count are, therefore,

1) . Choosing all the words in a document
2) Eliminating common function words by consulting a stop-words list
3) Computing the frequency of occurrence of all words in each document, and
4) Assigning the most frequently occurring terms as the index terms.

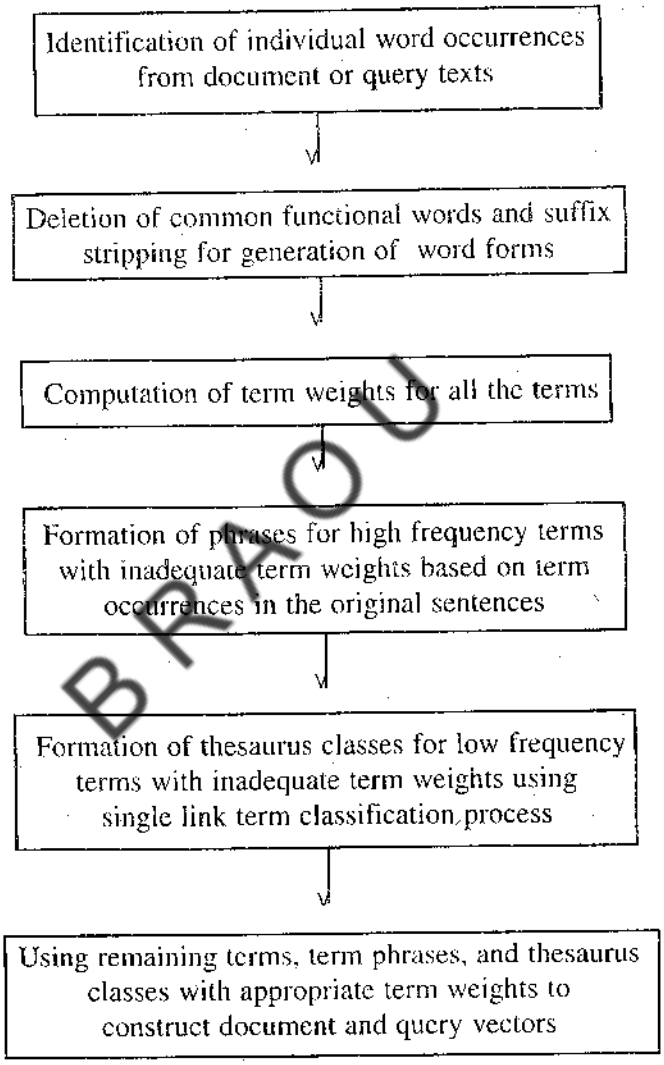
A close look to this method, which is based on the most frequently occurring terms, revel that such a system tends to produce better recall while neglecting the aspect of precision. This happens due to the underlying principle which says that if a term occurs several times in a document, then it should be regarded as significant one. However, it may be noted that all the documents in the collection on library science, say the term 'library' will occur quite frequently, but there is no need to use this term as the index term or a search term, because this will retrieve the whole collection. The objective of a good indexing system is to isolate all the documents in a collection from the others in the same collection which do not discuss the desired topic. In other words, one has to choose such words for indexing which can differentiate a given document or a group of documents from all the others in the same collection.

The steps for preparing an automatic index will be

1) Identification of all the words occurring in all the documents in a given collection,
2) Deletion of function words by consulting a stop-word list, and
3) Preparation of word stems by suffix-stripping method. (This method helps to decrease the number of words by producing a common stripped form for the words which have the same root, Eg: COMPUT for COMPUTING, COMPUTER, COMPUTATION, COMPUTATIONAL, etc.)

All approaches to the automatic indexing discussed above are mainly based on statistical or probabilistic calculations of term occurrences. Present-day working commercial indexing systems are all based on statistical analysis. Many information scientists, however, believe that in order to generate an effective index, the system has to consider the syntactic, semantic and pragmatic aspects of the words in the documents texts, rather than mere counting the occurrences of words in isolation.

SEMANTIC PRESENTATION OF AUTOMATIC INDEXING



Source: SMART Retrieval Project.

PROBLEMS INVOLVED

Compilation of a book index is relatively quite easy. But the construction of a bibliographical or subject index is not so easy. Finding out the correct subject of a document and specifying it in indexing language so that there is no problem of synonyms and homonyms, etc. But to retrieve a subject quickly on demand is rather a very complex job. In place of the ordinary or natural language, the indexing language is used to represent the subject contents

so that a searchable file is constructed for easy and quick retrieval. This is done because — i) there is a control of the use of synonyms and homonyms, ii) the vocabulary of indexing language is exact and precise, and iii) it is helpful in file organisation.

It is a rare phenomenon that the specific subject of a document is represented by a single term. Normally most of the documents, now-a-days, deal with compound and complex subjects comprising of a number of concepts. These subjects can be represented by more than one term. When a subject consists of a number of components, the ordering or sequencing of the components also brings in many problems.

Besides the problems of indexing, finding the location of the documents is equally difficult. It is not as easy as indicating the page number in the book index. The index entry represents subject contents and the location of the document is known through it. And a document can be a book, report, a dissertation, a thesis, a patent, a standard, an article in a composite book, or a microform. The reference for each one of these documents will have to be prepared according to definite cataloguing rules of catalogue code used for the relevant document. The rendering varies from document to document. This problem requires a good knowledge of practical cataloguing on the part of the indexer. The formulation of appropriate subject heading(s) representing the contents of a document has, thus, been a problem. Cutter, Kaiser and Coates have offered solutions to solve these problems. Chain procedure, PRECIS and POPSI are the other successful efforts towards this end.

6.4 DEVELOPMENTS IN SUBJECT CATALOGUING

A number of attempts were made in subject cataloguing right from Cutter's *Rules for Dictionary Cataloguing*. In this section we will discuss briefly the major contributions of Cutter, Kaiser, Coates, Ranganathan, Farradane, and Sharp.

6.4.1 Cutter's Contribution

Charles Ammi Cutter was the first to discuss the concept of specific subject in his *Rules for Dictionary Catalogue* in 1876. He advocated the entry of a work under its specific subject heading and not under the subject heading of the class which includes the subject. The 'stock subjects' under one of which Cutter wanted every book to be accommodated was not what we call a 'Specific Subject' today. Cutter also gave rules of entry for multi-word headings. He suggested entering "a compound subject name by its first word, inverting the phrase only where some other word is decidedly more significant or is often used alone with the same meaning is the whole name". Cutter, however, did not specify as how to find out which one of the component parts of a multi-word compound heading would be decidedly more significant. To find this out, the indexer was to rely on his own judgement or flair. This brought in some uncertainty in the application of this rule and fixing the order of various components in the subject heading.

6.4.2 Kaiser's Contribution

In 1911, J. Kaiser in his Systematic Indexing laid down a theoretical basis for fixing the order of significance among various components of a composite heading. He recommended

that a subject should be analysed to divide the constituent concepts into two categories: Concrete and Process. He categorised things, places and abstract terms not specifying action as Concretes and the terms signifying action as Processes. For instance, if the subject of document is the Classification of Books, then the books are Concrete and classification, a Process.

The order of components advocated by Kaiser was Concrete and Process, or Concrete was to be the entry word.

Locality-topic combinations, however, were to have double entry. Example: 'Export of Wheat' shall have WHEAT - Export as the subject heading. But 'Export of Wheat from India' shall have the following two subject headings:

WHEAT - India - Export
INDIA - Wheat - Export

Thus, by fixing the order of significance of the components of a compound subject, Kaiser endeavoured to resolve the uncertainty, which Cutter had left untouched. Though Kaiser did not solve all the problems yet his contribution towards the theory of indexing was fundamental.

6.4.3 Coates' Contribution

E.J.Coats' contribution towards subject indexing is very significant. His order of significance among various components of a compound subject is "Thing, Material and Action". According to Coates, "the most significant term in a compound subject is the one which is most readily available to the memory of the enquirer" or "the enquirer's choice which gives the most definite image". In a concept `Springing Cat', the image of cat comes to our mind first and then only we think of the act of springing. And without `cat' the idea of springing' cannot be visualised. The order of the component terms in a compound subject heading will have to be the same as our thinking process. Coates supports Kaiser's categorisation and the order of categories. His "concrete" and "process" have been renamed as "Thing" and "Action" by Coates. He develops his ideas further and introduces more categories such as "Material" and "Part" fixes their order of significance as follows:

"Thing", "Material", "Action"

"Material" is which the "Thing" is made of. It follows the "Thing" and precedes the "Action". Both "Thing" and "Material" produce static images, but the former produces a more definite mental image than the latter and, therefore, precedes in the order of significance. A "Thing" can have its parts, e.g., `nib' of a pen is directly related to it. A part will have to follow the "Thing" immediately and the order of significance is -

"Thing", "Part", "Material", "Action"

6.4.4 Ranganathan's Contribution

Early efforts to arrive at suitable subject heading were to some extent ad-hoc solutions without any sound theoretical base. The basic principles implied in the preferred sequences of composite headings are not very clear as alternatives were suggested without showing sufficient reasons in all cases. The clear understanding of the concept of specific subject and the vision to formulate it on scientific basis could only offer right solution to the problem.

Ranganathan did it through his Chain Indexing. It uses a classificatory base, and the indexer takes over from where the classifier leaves. The logic behind chain indexing in the words of Guha is "If an analytico-synthetic scheme of classification, having a structured notational system, is used in the coextensive representation of a subject then a retranslation of a class number, using the schedules of the same scheme of classification, would give a neatly structured formulation of the subject".

The components of a compound subject in chain indexing get automatically arranged in the order of Time, Space, Energy, Matter and Personality; or Personality, Matter, Energy, Space and Time.

6.4.5 Farradane's Contribution

The sequence of components in a subject formulation was all important before J.E.L. Farradane, who has provided a new approach which is distinct from others, creating a new type of syntax in indexing languages. His subject formulation is not based on the characteristics of the component terms. It is based on the relationship that exists between each pair of components. He recognises nine types of relationships, such as - association, comparison, concurrence, dimensional, distinctness, equivalence, etc. These relationships are indicated between two components to show their correct relationships. Operator ';' indicates 'association'. The subject formulation for the document, "Diseases of eyes" shall be Eyes/;Diseases. Similarly, the subject formulation for a document, "Treatment of the Diseases of Eyes" shall be Eyes/ ;Diseases/-Treatment. Operator '-' represents 'Action' relationship here. This subject formulation can also be represented in the following order of components:

Treatment-/ Diseases;/ Eyes.

Both the components and operators have been reversed. By changing the direction of the operators, the same relationship can be retained between pairs of terms. This naturally lessens the importance of sequence or order of significance of terms in a subject formulation.

6.4.6 Sharp's Contribution

In conventional indexing, access is provided from each of the component terms by rotation or cycling so as to have a network of references. It is done to ensure maximum access. Against this, Sharp is of the view that specific subject formulations, having all the components, may not be of much use to the seekers of information. He believes that most of the users are able to cite more than one but not all the components constituting a specific subject. Also the users do not get correct answers because the combination of components they search for are not there. Keeping these points in view, Sharp in his SLIC Indexing System ensures optimum economy in the bulk of index by providing combinations irrespective of the citation order. Sharp says that cycling by complete permutation will not do. The method he suggests is combinations, that is, selection of groups. Sharp finds out only some selective combinations of components constituting the subject of the document for using as index headings.

6.4.7 Lists of Subject Headings

Subject indexing cannot be practised by the indexer in his own way as the entries in subject index are made in accordance with specific rules governing their choice, style and

terminology. If compiled otherwise, the subject index is bound to have many variations in subject headings which in turn render it less useful to the user.

Making of author or title entry is easier than making of a subject entry, because both author and title are established facts and are available from the title page of the document itself. Then rules are also available for making the author or title entries. But the heading to represent the subject of the document coextensively is to be the indexer's decision and hence, the need of proficiency and maturity in the indexer's knowledge of the subject and his approach to the document. The efficiency of an index depends on the appropriateness of the chosen subject heading by the indexer. Cutter's Rules for Dictionary Catalogue or The Vatican Rules may be of some help in formulating an appropriate subject heading which may be helpful in retrieving the document when asked for.

Determination of subject is not easy. Some of the difficulties encountered in this arduous job are: i) subjects, most of the times, do not exist in separate compartment, ii) their scope is not constant, iii) they have no specific names, iv) they cannot be represented by a single term, v) they may have more than one name, vi) a document may have more than one subject, etc. To derive a subject heading, one method is to make use of a printed list of subject headings. Following are popular, important and well-known lists of subject headings:

1) Sears List of Subject Headings (SLSH)
2) Library of Congress List of Subject Headings (LCSH)
3) Medical Subject Headings (MeSH)

These lists aim to provide one form of subject headings under which all documents on a given subject can be listed.

6.5 INDEXING LANGUAGES

Once information is ascertained, it is to be recorded in some indexing language. In this section we shall discuss about the types of indexing languagges, natural language and artificial language.

6.5.1 Derived-Term Type and Assigned-Term Type

Indexing language is of two types: Derived-Term Type and Assigned-Term Type. In the Derived-Term system, all index terms are taken from the document itself. In the Assigned-Term system, it is the indexer who constructs the index terms or descriptions. The latter is an intellectual method involving the finding out of specific subject of the document and assigning an appropriate subject heading. Thus, author indexes, title indexes, citation indexes and natural language indexes are derived term systems, whereas all indexing languages with vocabulary control devices, such as subject heading lists, thesauri and classification schemes are assigned term systems. The derived term systems are almost clerical and can be easily mechanised. The assigned term systems, on the other hand, are intellectual and, therefore, require more time and money at the input stage. A lack of structuring, logic or sense in indexing will produce irrelevant output.

6.5.2 Natural Language

Natural language has some advantages. Its vocabulary is uptodate and it keeps on growing assimilating new concepts as soon as they come into being. It has syntax and rules of grammar which enable it to convey correct meaning of a specific subject. These advantages not withstanding, natural language is not helpful in organising an index file which is a must for information retrieval. The natural language suffers from the problem of homonyms and synonyms because of its flexibility.

6.5.3 Artificial Language

The lists of subject headings, classification schemes and thesauri are representatives of indexing languages. Because of the controlled vocabulary of indexing languages, it is exact and precise. And there is no problem of synonyms and homonyms because 'see' and 'see also' references as also 'use' references take care of these. The micro-documents of these days require a number of components to specify their subject contents. The sequencing of these components is done according to the rules of syntax of a particular indexing technique. For instance, the Chain Procedure, PRECIS, POPSI, etc. have their own rules of syntax for the formulation of subject representation of a document.

6.6 CHAIN PROCEDURE AS SUBJECT INDEXING

Chain Procedure is a semi-mechanical method to derive subject index entries or subject headings from the class number of a document. Ranganathan's idea about the role of classification was to analyse the subject of a document into its fundamental components and then to synthesize these components in a logical order in the classificatory language. The resulting class number was to be co-extensively expressive of the specific subject of the document. This class number is to be the basis of chain indexing.

An indexer is not supposed to analyse the subject of the document. This is the job of a classifier. The indexer is supposed to start from where the classifier has left. No duplication of work is to be done. He is to draw subject headings or class index entries from the class number of the document to provide alphabetical approach to the subject of the document. Cataloguing, in fact, is to supplement the work of classification through alphabetical index entries derived mechanically by chain procedure. Ranganathan calls it the 'Symbiosis' of classification and cataloguing.

6.6.1 Steps in Chain Procedure

According to Bhattacharya, there are as many as eleven steps involved in chain procedure.

1) Determination of the specific subject of the document is done with the help of the title of the document, its table of contents and by a careful perusal of the text. By analysing the subject contents of a document, one arrives at its specific subject.

2) Naming of the specific subject of the document expressively in the natural language in terms as decided (expressive name-of-subject).

3) Representation of the name of the specific subject in terms of its fundamental components (name-of-subject in kernel terms). It is done by removing all the auxiliary words from the title.

4) Determination of the category or status or role of each fundamental component according to a set of postulates and principles formulated for this purpose (analysed name-of-subject).

5) Transformation of the analysed name-of-subject by rearranging, if necessary, the fundamental components, according to a few additional postulates and principles formulated for the purpose of governing the syntax (transformed name-of-subject).

6) Standardization of each term in the transformed name-of-subject (name-of-subject in standard terms). If the name-of-subject is not in accordance with the standard terms used in preferred scheme of classification, it should be replaced by its equivalent standard terms, as given in the schedule. If the terms in the schedule are not current, help of a thesaurus or a glossary of the subject may be taken.

7) Determination of each of the links of the chain in which the subject denoted by the name-of-subject-in-standard-terms is the last link (determination of under link). Representation of the class number in the form of a chain in which each link consists of two parts - the class number and its translation. It is done as follows:

   a) Make the first link from the first digit

   b) Make the second link out of two digits and so on, upto the last link (ie link occurring last in the chain produced by a class number) which is to be made of all digits

   c) Write the links one below the other in succession

   d) Write against each link its translation into natural language

   e) Connect each link with its translation by an "=" sign, and

   f) Join the "=" sign of each link with that of the next succeeding link by a downward arrow, if necessary.

8) Determination of the different kinds of links, such as, Sought Link, Unsought Link, False Link and Missing Link (determination of kinds of links). False Link (FL) is that which is not a Class Number. It does not represent a subject with a definite name. A link is FL if it ends with a connecting symbol or digit representing phase relation, or time isolate idea representing time itself. Unsought Link (UL) is that which ends with a part of the isolate focus in a class number, or represent a subject on which reading material is not likely to be produced or sought or which is not likely to be looked up by any reader seeking material on the specific subject. Missing Link (ML) is a link in a chain with gap corresponding to the missing isolate in the Chain. Sought Link (SL) is that link through which a user approaches his document. It is neither a False Link nor an Unsought Link.

9) Derivation of subject heading from each of the sought links in the Chain according to a set of rules formulated to suit the purpose at hand (derivation of subject headings). The procedure for deriving subject headings is to start from the terms of the last Sought Link and proceed towards the terms of upper link, in a reverse rendering process.

10) Construction of specific heading for specific subject entry or subject reference entry is to be made with the minimum number of terms of such upper links as are necessary and sufficient to make the subject heading meaningful and individualised. Each term in the heading or sub-heading is to be a single noun in nominative case except when a qualifying adjective is necessary as in 'Derived System' or 'Social Sciences'.

11) The specific subject entries, subject reference entries and entries for cross references should be merged and arranged in a single alphabetical sequence.

6.6.2 Chain Indexing in Operation

With the above background, the actual operation of Chain Indexing can be further studied with the help of some examples. To recapitulate, it may be stated that Chain Indexing is a procedure for deriving alphabetical subject headings (class index entries) through the digit by digit interpretation of the class number of a document. A class number generates a series of links forming a chain.

Example-1:

The document entitled, Mr.Whitman, by Stanley T.Pullen, having the class number O111,1M56,1 will generate the following chain:

O	=	Literature (SL)
O1	=	Teutonic Literature (UL)
O11	=	Indo-European Literature (UL)
O111	=	English Literature (SL)
O111,	=	FL
O111,1	=	Poetry, English Literature (SL)
O111,1M56	=	Whitman, Poetry, English Literature (SL)
O111,1M56,	=	FL
O111,1M56,1	=	Mr.Whitman, by PULLEN (Stanley T) (SL)

In the above chain, O = Literature is known as the upper link and the lower links is O111,1M56,1 = Mr.Whitman, by PULLEN (Stanley T). The links are represented by the digits. Connecting digits, however, do not give any meaningful verbal interpretation. Therefore, these are termed as False Links. O111, and O111,1M56, are False Links in the above Chain. Two links O=Teutonic Literature and O111 = Indo-European Literature are Unsought Links because they represent subjects which are not likely to be sought by the users and under the present context there may not be documents under these subjects. Rest of the links in the above chain are Sought Links. They convey meanings directly related to the document under reference and the users are likely to approach this document through them.

Corresponding to these five Sought Links, the following subject headings or class index entries will be generated by the above Chain:

Mr.Whitman, PULLEN (Stanley T) = O111,1M56,1

PULLEN (Stanley T), Poetry, English = O111,1M56

Poetry, English, Literature = O111,1

English Literature = O111

Literature = O

The lower link in the above chain represents the specific subject of the document, rest of the links are of greater extension.

Example-2

The document entitled "Paraffins" with class number 547.411 generates the following Chain:

500	=	Science (UL)
540	=	Chemistry (SL)
547	=	Organic Chemistry (SL)
547.	=	FL
547.4	=	Alphatic Compounds, Organic Chemistry (SL)
547.41	=	Hydrocarbons, Organic Chemistry (SL)
547.411	=	Paraffins, Organic Chemistry (SL)

The above Chain has seven links in all. Out of these, one is a False Link, two are Unsought Links and four are Sought Links. Corresponding to the Sought Links, there are four subject headings or Class Index Entries as follows:

Paraffins, Organic Chemistry	=	547.411
Hydrocarbons, Organic Chemistry	=	547.41
Alphatic Compounds, Organic Chemistry	=	547.4
Organic Chemistry	=	547

6.6.3 Merits of Chain Indexing

Some of the merits of the Chain Indexing system are as follows:

1) The classifier analyses the subject of the document to have a structural formulation of the subject as its class number. The indexer starts from this stage and retranslates the class number to provide alphabetical approach through class index entries. Chain indexing, thus, saves the duplication of work.

2) Chain Indexing is based on the classification number as well as the terminology, given in the schedules. The indexer has to judge which term is to be taken as sought link and which one to be omitted as unsought link. It is, therefore, semi-mechanical as also speedy procedure.

3) Subject headings or class index entries can be derived from class number of any scheme of classification system with the help of chain procedure. Mills has also shown that chain indexing can be applied with ease to any classification scheme whose notation symbols indicate the subordination of each step of division.

4) For a string indexing four components, only four subject headings are made according to chain indexing, though the permutation of four terms would have given 24 headings. This system thus brings in tremendous economy.

5) Chain indexing provides alternative approach to the classified file through reverse rendering. It is helpful to the users in retrieving their information.

6) Documentation Research and Training Centre (DRTC) has found the chain procedure fully amenable to computerisation. Programmes were successfully written to generate subject headings both from class numbers and feature headings following a conventional reverse rendering method.

7) Chain Procedure may also be used to derive indexes to classification schemes and books. It may be used in formulating headings necessary for guide cards in a catalogue, stack rooms guides, gangway guides, bay guides, shelf guides, etc in consistent way.

6.6.4 Limitations of Chain Indexing

Chain indexing system has some limitations too. Some of them are given below:

1) Out of the subject headings generated for a document through Chain Indexing, only the last one is specific and others represent broader subjects. The specific subject heading shall be available to only those who have a particular search formulation.

2) Besides the generic entries, Chain Indexing may also generate some entries for empty links in the chain in the case of a document of a highly specialised field. These empty links create noise problem in the file.

3) As Chain Indexing uses the class number for drawing subject headings, its efficiency or otherwise depends on the scheme of classification. A procedure dependent on the scheme of classification has to share the defects of that scheme.

6.7 LET US SUM UP

Subject Indexing is a method of information retrieval. Due to complexity in the nature of development of subjects and the way information seekers go about it, access to relevant subject at any time needs efficiency in recall. Subject Indexing, in general, leads to access through term control. A searcher expresses in the form of a rough statement sometimes complete but mostly incomplete. The subject index should help leading a searcher from a vague statement to expressive statement. The system of control of searching should be such that a searcher is helped to arrive at a statement compatible with the information base that is being searched. This is done by either broadening or narrowing down the concepts given in a statement of the searchers. Thus, a subject indexing system affects a kind of compatibility between the searcher for an information and the information base.

The refinement of indexing techniques to provide better performance is a continuous process. Melvil Dewey, C.A.Cutter, J.Kaiser, S.R.Ranganathan, E.J.Coates, J.E.L.Farradane and many others have contributed significantly towards the development of indexes and indexing. The indexer summarises the contents of the whole document in a few words. The subject analysis means the identification and express of subject matter in document texts, databases, etc.

Most of the documents deal with compound and complex subjects, each comprising of a number components or concepts. The coordination of these components is either done at the input stage or at the output stage. The index in which coordination of components is done at the input stage, is known as Pre-Coordinate Index; and where coordination is effected at the output stage, it is termed as Post-Coordinate Index.

The search strategies, process, both manual and automatic, including their limitations are discussed in this unit. The indexing languages and the techniques are explained while emphasising the chain index and its impact on subject indexing.

6.8 REFERENCES AND FURTHER READING

BHATTACHARYYA, G. "Postulate-based permuted subject indexing system". *Library Science.* 16; 1979. (Paper A)

BHATTACHARYYA, G and A.Neelameghan. "Postulate-based subject headings for dictionary catalogue system". *DRTC Annual Seminar;* 1969. (Paper CA)

BROWN, A.G. *Introduction to subject indexing,* 2nd ed. London: Clive Bingley, 1982.

CHAKRABORTHY, A.R. and B. Chakraborti. *Indexing: Principles, processes and products.* Calcutta: World Press, 1984.

CHAN, Louis Mai. *Library of Congress subject headings,* 7th ed. Colorado: Clive Bingley, 1982.

CHOUDHARY, G.G. *Information retrieval systems.* Calcutta: IASLIC, 1993.

COATES, E.J. *Subject catalogues.* London: LA, 1988.

FOSKETT, A.C. *The subject approach to information,* 4th ed. London: Bingley, 1982.

GOSH, S.B. and J.N.Satpathi. *Subject indexing: Concepts, methods and techniques.* Calcutta: IASLIC, 1997.

INDEXING Systems! edited by T.N.Rajan. Calcutta: IASLIC, 1981.

LANCASTER, F.W. *Information retrieval systems: Characteristics, testing and evaluation.* 2nd ed. New York: Wiley, 1979.

LANCASTER, F.W. and L.C.Smith. *Compatibility issues affecting information systems* and services. Paris: Unesco, 1983.

PRASHER, R.G. *Index and indexing systems.* New Delhi: Medallion, 1989.

ROWLEY, Jennifer E. *Organising knowledge: an introduction to information retrieval.* Aldershot: Gower, 1987.

VICKERY, B.C. *Techniques of information retrieval.* London: Butterworths, 1970.

6.9 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) What is Subject Indexing ? Explain the basic steps involved in indexing process.

2) Explain the advantages of automatic indexing over manual indexing.

3) What is Chain Procedure ? Discuss its merits and demerits with suitable examples.

II SHORT NOTES

a) SLIC

b) Kaiser's Systematic Indexing

c) Exhaustivity *vs.* Specificity in Indexing

UNIT - 7 : PRECIS AND POPSI

Structure

7.0 AIMS AND OBJECTIVES

The unit aims to introduce you to the two most important pre-coordinate indexing systems, namely, PREserved Context Indexing System (PRECIS) and Postulate-based Permuted Subject Indexing (POPSI).

After studying this unit, you will be in a position to

- describe briefly the development of PRECIS and POPSI
- explain the main features and structure of PRECIS and POPSI

- list out the advantages and limitations of PRECIS and POPSI

- make a comparative study of PRECIS and POPSI and assess their effectiveness in subject indexing.

7.1 INTRODUCTION

An index is a systematic guide to the intellectual content and physical location of knowledge record. It is a key to open references to an item in the text of document or in a collection of documents. Basically, an index has remained as a tool to help a user to retrieve his information bearing documents from a collection. Any information system exists to provide the seeker of information any document which bears his information or answers his query.

An index is an operational tool that helps the information system to achieve his goal. A library catalogue is a typical example of an index which helps the user to reach near his documents which may contain his information. Usually, an index is arranged alphabetically, but sometimes may be arranged chronologically, geographically, numerically or in any other suitable manner, depending upon its requirements.

The term 'Index' has been derived from the Latin word 'indicare' which means 'to indicate' or 'to point out'. Index has been defined differently by different authors. In *Encyclopedia of Library and Information Science,* J.Rothaman defines an index as "a pointer or an indicator, more often alphabetic that includes subjects and names of people and places that are considered to be of special significance in a graphic record". *The British Standard BS3700:1954* defines an index as a systematic guide to the location of words, concepts or other items in books, periodicals or other publications. An index consists of series of entries appearing, but in the order in which they appear in the publication, but in some other order (eg., alphabetical) chosen to enable the user to find these quickly, together with references to show where each item is located.

In simple terms 'Indexing' means the art and science of preparing an index. As an art indexing appreciates the use of sense and taste and as science it requires the use of rules and patterns. Brenner (1979) defined indexing as "the process of analysing the information content in the language of the indexing system". UNISIST (1975) regarded indexing system as "the fact of describing and identifying a document in terms of its subject content". According to Lancaster (1979), subject indexing involves two quite distinct intellectual steps — conceptual analysis or content analysis of a document, followed by the transformation of the conceptual analysis into "index language". With the growth of published literature both in quantity and complexity and with the realisation of the importance of information in research and decision making, the value and importance of an index grew steadily.

Over the ages, there has been considerable improvement in the quality and design of the index or indexing services to meet the variegated requirements of the users and also to act as an effective communication link between the source of information and the user of information. The producers of information generate new information for communication and use. The librarians and information scientists organise this information for systematic storage and quick and efficient retrieval. It is primarily for the users that the different type of the indexes and the indexing techniques have been developed.

By nature, the alphabetical indexing methods are basically co-ordinate indexing where the concepts of the subject headings referring different documents are co-ordinated. Thus, co-ordination in the context of 'Indexing', therefore, means the combination of two or more concepts to create a new concept. The purpose of co-ordination of concepts of subject headings of documents is to describe the contents of the documents indexed more accurately. In this sense co-ordinate indexing is in essence "Concept co-ordination".

The accurate content representation (Document representation) in the subject headings increases the specificity and results in minimising the retrieval of irrelevant documents during a search. PRECIS and POPSI are regarded as pre-coordinate indexing systems besides chain indexing. In the pre-coordinate indexing systems, the coordination of index terms is done at input stage in anticipation of the user's approach. In these indexing systems the subject of a document is analysed into constituent concepts and the concepts are then represented by symbols or words used in the indexing language (the language of the index). These symbols or words are finally arranged in a logical order following the system of the indexing language to derive the subject formulations. All these pre-coordinate subject formulations are maintained in an order in the index file as a key to the contents of the document collection. The main characteristics of pre-coordinate indexing systems are as follows:

i) They are based on subject analysis of documents

ii) They follow some specified citation order of terms, and

iii) Co-ordination of terms is done at input stage.

Since searching is basically a matching operation, the success of an indexing system depends on the ways and the constituent index terms are pre-coordinated that will match the approach of the majority of the searchers. Each indexing system, therefore, introduced its indexing language. Each system has framed its own "rules of syntax" to achieve the most preferred order of coordination of index terms. Besides, determining the preferred citation order, the rules also provide guidelines for expressing relationships between index terms.

In pre-coordinate indexing system both the indexer and the searcher are required to understand the mechanism of the system. It is necessary for the indexer in order to arrive at the most preferred citation order. It is essential for the searcher to formulate an appropriate search strategy to achieve the highest possible degree of matching concepts. The searcher has no other choice but to try to predict the citation order specified by the indexer. Thus, the pre-coordinate indexing system is obsessed with the consideration of best citation order often based on psychological investigations into the subject formulations in the reader's mind.

7.2 PRESERVED CONTEXT INDEX SYSTEM (PRECIS)

PRECIS is an acronym which stands for PREserved Context Index System. It is essentially a system for producing alphabetical subject indexes in a page format including paper, microform, or display on a computer terminal. A PRECIS index is, generally though not invariably, produced by a computer.

The labour involved in index production is divided between the human indexer and the computer. The indexer undertakes the intellectual tasks which require human judgements, i.e.,

determining subject content of the document, producing an input string, establishing indexing terms with references and issuing instructions for the computer. The computer, on the other hand, follows these instructions and carries out the mechanical chore of implementing the human decisions, producing the correct entries and references. The PRECIS is generally presented as a two-stage index, i.e. a subject index in which each entry is followed by one or more addresses (DDC class numbers in the case of BNB), which indicate the positions of the relevant bibliographic entries in another file.

The main objectives of formulating this semi-automatic method of indexing were as follows:

i) All the entries produced by this method should be co-extensive with the subject of the document

ii) Each of the entries should be meaningful and allow the users to interpret the entry correctly

iii) The use of computer for generation of computer entries

iv) There should be adequate system of references

v) The system should be able to admit freely new terms into the system as soon as they are encountered in literature

vi) There should be common set of indexing rules in order to have consistency of work.

7.2.1 Brief History and Development of PRECIS

PRECIS was developed and used by the British National Bibliography (BNB) for producing entries in the subject index in the weekly and cumulative issues of the bibliography which lists new British books.

The BNB became involved in the UK-MARC Project in 1968. It was then contemplated that a new indexing system capable of providing co-extensive subject indexing for each document in the MARC database and amenable to computer manipulation was needed. Under the direction of Derek Austin, a project was undertaken to develop such a system. After initial experiments and trials a prototype version of a new indexing system called PRECIS was adopted by BNB from January 1971 till the end of 1973.

In January 1974, a new and improved version of PRECIS became operational and this system is being used by BNB and a number of other bibliographies, indices and library catalogues, mostly in the British Commonwealth countries. In 1984 the old 1974 version of PRECIS is revised and updated to meet a number of developments over the past ten years.

Recent developments and research on the system include a translingual project, studying the applicability of PRECIS in languages other than English and the feasibility of automatic language switching by the help of a multilingual thesaurus designed for computer manipulation.

In developing PRECIS, Austin applied the modern theory of classification as well as linguistic principles. Many of the theories and ideas developed in the 1960s by the Classification Research Group (CRG) in Britain in search of a general classification scheme were adopted in the development of PRECIS by Austin, who has been an active member of the group.

When PRECIS began, it was based on modern classification theory, but it has moved gradually towards a linguistic analysis approach. Breaking down a subject into its component parts is based on classification theory. Recording the parts into a meaningful string, draws upon linguistic principles. The relationships shown in the thesaurus are based on classification principles.

One basic principle of PRECIS is that each entry should represent the complete theme or topic of a document in summary form. This is called the Principle of 'Co-extensivity'.

The basic criteria for useful index entries have been outlined as follows:

1) An entry can be made under any term likely to be sought in a string

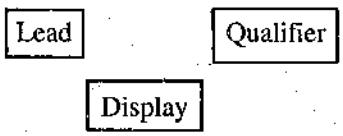
2) Each entry should be intelligible, and it should state the subject clearly

3) Entries should be consistent in structure, so that they collocate with those produced from other strings or similar themes.

All these entries lead to the development of the principle of 'context dependency'. This principle requires that the individual concepts in an entry should be organised in a one-to-one relationship in context-dependent order. In other words, each term in the entry is related to one immediately preceding and the one immediately following it. Thus, each term sets the next term into its obvious context.

7.2.2 Features/Structure of PRECIS

In order to preserve the one-to-one relationships in context-dependent order, each PRECIS entry is presented in a two-line and three-part format. This is the standard format of PRECIS which is show below:



Lead is the term which acts as out approach term.

Qualifier is that term or set of terms which qualifies the Lead term to bring it into its proper context. It provides wider context of the Lead term. Display is the remaining part of the string which helps to preserve the context. These terms are context-dependent on the lead term.

Following is an example of a PRECIS input string with the entries generated from it.

Input string:

(0) France
(1) textile industries
(p) skilled personnel
(2) training

Entries:

France
Textile industries. Skilled personnel. Training

Textile industries. France
Skilled personnel. Training

Skilled Personnel. Textile industries. France
Training

Training. Skilled personnel. Textile industries. France

A good index provides all possible approaches to the document so that there is no difficulty in its retrieval. In order to achieve this object, every component term in a string should serve as our approach term, by turn, for the benefit of users. At the same time all the entries so derived should be able to specify the clear context of the document.

In the aforesaid entries, a separate entry is generated with each significant term on the lead position. The lead which contains the entry element, printed in bold typeface must be occupied. The qualifier may contain any number of terms, which are of successfully wider context to the right of the line. The terms in the display are of progressively narrower context to the right of the line.

The device of bringing each significant term to the lead position while maintaining the context-dependent order is called ***Shunting***. The two-line and three-part format and the established procedure of shunting ensure that each concept is placed immediately next to those most closely related to it, whatever position this particular concept may occupy in a particular entry.

In the second entry, in the example above, the concept "textile industries", when it is moved to the lead position remains next to both "France" and "skilled personnel".

Inverted Format

Besides the standard format there is an inverted format. This format is used when any term prefixed with role operators (4), (5) and (6) becomes a lead term.

For example, in the subject "Economic aspects of nuclear power" the input string and index entries will be as follows:

Input string:

(2) nuclear power
(4) economic aspects

Entries:

Nuclear power
- Economic aspects

Economic aspects
Nuclear power

The following characteristics of the above entries should be noted:

(a) The term coded '4' is printed in italic following a long dash whenever it occurs in the display. The long dash is also retained when the 'view point' term appears as the first component of the display.

(b) The term coded '4' is printed in bold when it appears in the lead. The entry is then produced in the inverted format: the display consists of terms selected from the string in their input order.

(c) The term coded '4' is dropped from the display when it appeared in the lead. It would have been repeated in the display in any one of the following circumstance:

i) if the lead did not contain the whole of the term. (For eg: if only the focus had been marked as lead);

ii) if the string contained a later term (coded '5' or '6')

iii) if the term formed part of a block, and the whole of the block was not present in the heading (i.e., the lead + qualifier). This applied particularly to term coded '5' which always occurs as the first components of blocks.

Predicate Transformation Format

The Predicate Transformation Format is used when the term representing performer (3) appears as a lead term prefixed by one of the operators 2,s,t,&,u. The following input string and the index entries generated by it will make this format clear.

Subject: Hunting of rodents by foxes

Input string: (1) rodents
(2) hunting $ v by $ w of
(3) foxes

Entries: **Rodents**
Hunting by foxes

Hunting. Rodents
By foxes

Foxes
Hunting of rodents

Entries under concepts with roles as performers are generated in stages as follows:

Stage - 1 : The performer term is assigned to the lead

Stage - 2 : Since a performer term is prefixed by operator'3', the computer then automatically reads the operator assigned to the next preceding term, checking for the presence of an action. Action concepts are identified by one of the four operators, namely, '2', 'u', 's' and 't'

Stage - 3 : If an action term is present, it is assigned to the display, i.e.,

Foxes
Hunting

Stage - 4 : If the action is also accompanied by an up-ward-reading connective ($w), the computer continues producing the phrase in the display, eg.,

Foxes
Hunting of rodents

Stage - 5 : When the string contains no further candidates for the phrase in the display (ie., terms accompanied by $w), the remaining terms in the string (if any) are assigned to their standard format position.

Example: *Training of apprentices by foremen in the aerospace industries*

Input string:

(1) aerospace industries
(p) apprentices
(2) training $v by $w of
(3) foremen

Entries:

Aerospace industries
Apprentices. Training by foreman

Apprentices. Aerospace industries
Training by foremen

Training. Apprentices. Aerospace industries
By foremen

Foremen. Aerospace industries
Training of apprentices

Two-Part Structure: Syntax and Semantics

In PRECIS, the two categories of relationships, between terms used in indexing are recognised and handled by different procedures, namely, syntactic relationships and semantic relationships.

Synthetic relationships refer to *a posteriori* relationships which are document-dependent. In other words, the terms are not originally related but the relationship has been established within the context of a particular document. For example, a work about hotel management establishes a relationship between the concepts "hotel" and "management" which possess no inherent relationship outside the context of a document. Such relationships refer to the organisation of terms in input strings and their manipulation into entries. Since the order of the terms in strings is regulated by a kind of `grammar' we might call this the syntactical side of the system. Semantic relationships on the other hand refer to a priori relationships which are invariable and independent of the treatment of the concepts in any particular document. The two main types of semantic relationships are hierarchy and synonymy.

In PRECIS semantic relationships are handled in a thesaurus and such thesauri relationships are maintained between indexing terms and their synonymous broader terms, narrower terms, etc. The syntax and semantics are handled by two separate procedures in PRECIS which are discussed below.

Syntax : This procedure deals with the analysis of individual documents and the assignment of input strings. The first step is to identify the subject of the document and the next step is to separate the individual concepts in the subject and identify the relationships between them.

In order to maintain the one-to-one relationships in context dependent order and to achieve consistency in indexing practice, a scheme of role operators (Figure-1) has been developed which served as the indexer's grammar.

Each term in a string is coded with a role operator which expresses, in a machine readable form, its role in the subject and its position in the string. The role-operators also have built-in computer instructions with regard to format of the index entry, typography of each term and its associated punctuation.

A string assigned to the subject of a document may contain any number of terms. But each string must begin with a primary operator in the range from 0 to 2 and must contain at least one term coded either (1) or (2). There is a direct connection between the primary operators 0,1,2 and 3 and contain grammatical structures in everyday speech. The operator 0 corresponds to the locative case in grammar, The operators 1,2 and 3 are used with terms which generally correspond to the object, verb (in the case of transitive verb) and subject of a sentence.

In the case of intransitive verb, the operators 1 and 2 are used with terms which function as the subject and verb of a sentence. The role operators are devices for constructing the input strings and serve as instructions to computer. They do not appear in the index entries.

Role Operators of PRECIS

Main Line Operators	0	Location
Environment of Observed system	1	Key system: object of transitive action; agent of intransitive action
Observed system (core operators)	2	Action/Effect
	3	Agent of transitive action; Aspects : Factors
Data relating to observer	4	Viewpoint-as-form
Selected instance	5	Sample population/Study region
Presentation of data	6	Target/Form
Interposed operators		
Depended elements	p	Part/Property
	q	Member of quasi-generic group
	r	Aggregate
Concept interlinks	s	Role definer
	t	Author attributed association
Coordinated concepts	g	Coordinated concept
Differencing operators (prefixed by $)	h	Non-lead direct difference
	i	Lead direct difference
	j	Salient difference
	k	Non-lead indirect difference
	m	Lead indirect difference
	n	Non-lead parenthetical difference
	o	Lead parenthetical difference
	d	Date as a difference
Connectives (Components of linking phrases; prefixed by $)	v	Downward reading component
	w	Upward reading component
Theme interlinks	x	First element in coordinate theme
	y	Subsequent element in coordinate theme
	z	Element of common theme

Source: GUHA, B. Documentation and Information: Services, techniques and systems, 2nd rev ed. Calcutta: The World Press, 1983.

Example of Use of Role Operators :

Subject : *Car production in Germany*

String :

(0) Germany
(1) Cars
(2) production

Entries :

Germany
Cars. Production

Cars. Germany
Production

Dependent Elements

The role operators (p), (q) and (r) are dependent elements. These follow immediately the term on which they are dependent.

Example:

(1) aircraft
(p) noise

Compound Terms

Many concepts are represented by compound terms. In most cases it is desirable to bring each of the elements into the lead. A device called differencing is used to identify each component of a compound term. The word or (words) coded as a difference is a modifier. If it modifies the noun or the focus of the term, it is called direct difference and coded as $01 if it is not to appear as a lead, or with $21 if it is used as a lead. If the word modifies another word (or words), and is coded as a difference itself, it is called an indirect difference and coded with $ 02 if nonlead and $22 if lead.

Example : *Subject reinforced concrete bridge*

String :

(1) bridges $ 21 concrete $ 22 reinforced

In the said example "concrete" is a direct difference and "Reinforced" is an indirect difference.

Connectives and Substitutes

In many cases, in order to preserve natural word order, the device of connectives and substitutes are needed.

The role operator $ v indicates a down - ward - reading connective, and $ w an upward - reading connective.

Subject : *Application of computer systems in design of electronic equipment*

String :

(1) electronic equipment
(2) design $ w of
(s) application $ v of $ w in
(3) computer systems

Entries ;

Electronic equipment
Design. Application of computer systems

Computer systems
Application in design of electronic equipment

Semantics

In order to achieve vocabulary control all terms which have been used as lead terms in indexing are entered in a machine - held thesaurus. In addition non-preferred synonymous terms, other related terms are also included in the thesaurus because these also function as user's access points. Relationships between each term and other terms are indicated.

There are three basic thesauri relationships : equivalence, hierarchical, and associative.

1. Equivalence relationship - code ($m)

   a. Synonymous
      Birds / Aves

   b. Quasi synonymous
      Hardness / Softness

2. Hierarchical relationship - code ($0)

   a. Generic relationship

      Rodents
      Mice
      Rats

   b. Hierarchical whole - part relationship

      Geographical regions
      United States
      California
      Los Angels
      San Francisco

3. Associative relationship - code ($n)

This relationship exists between terms which are not synonymous nor hierarchically related, but are nevertheless mentally associated. One of the terms is entailed by the others, and frequently plays a part in its explanation or definition.

Birds / Ornithology

Reference link the related terms. See references are made from terms which are not used in index entries to referred terms which are used consistently to represent that concept. See references are used in equivalent relationships See also references are made between terms either of which could appear as lead in an index entry. The codes ($n, $0, and $n) expressing thesauri relationships have built-in machine instructions for generating appropriate references for a particular index. for hierarchically related terms See also references are made from the broader term to the narrower, but not reciprocally. See also references are also used for Associated relationships.

The references are generated from a computerised file which consists of independent but interrelated addresses.

The codes have built-in machine instructions for generating all references indicated. However, if all the terms in a network of a particular index have not been used as lead, there is bypass routine for avoiding blind references i.e., references to terms under which there is no entry in that index.

7.2.3 Application of PRECIS

In the first place the indexer has to determine the subject matter of the document. Usually, the subject matter is to be formulated in the form of a title - like purchase , such as; The training of skilled personnel in the Indian textile industries. In the next step the syntactical roles of the component concepts are to be determined and the relevant role operators which express these roles are to be assigned to the concept terms.

In the determination of the syntactical roles, the indexer would do well to ask first whether or not a term which denotes an action is present. If present, the action will usually determine how the remaining terms would be handled, just as the verb tends to dominate the sentence in the grammar of a natural language. In the phrase under consideration, it can be easily recognised that the term 'training' denotes an action. This term should therefore be prefixed by the role operator (2), as under :

(2) training

The next thing the indexer has to look for is the kind of action represented by the term, that is, transitive or intransitive action. In this case, 'training' represents a transitive action as it is capable of taking an object. When this is the case, the indexer is to look for the concept which is semantically related as the object of transitive action. In our example, it is the 'skilled personnel', who are being trained. This concept is usually coded as the key system (role operator 1).

When the above analysis is checked in the context of the other concepts present in the subject, it would be evident that 'skilled personnel' is actually part of some other named system, namely, 'textile industries'. To indicate this whole-part relationship, the role operator 'p' can be used. The concept 'textile industries' can be considered as the key system and 'skilled personnel' can be taken as part of the key system. Hence, the relationship between the three component concepts considered so far may be shown with their respective roles as :

(1) textile industries
(p) skilled personnel
(2) training

The remaining concept in the subject, namely, 'India' clearly functions as the environment in which the author considered all the other phenomena of his study. This should therefore be introduced as the location (role operator 0), which gives us the final string as :

(0) India
(1) Textile industries
(p) Skilled personnel
(2) Training

It would be observed that the number operators are arranged according to their ordinal value, while a non-numerical operator is attached to the concept with which it is related. In the above example, (p) skilled personnel has been attached to (1) textile industries, whose part it is according to our syntactical analysis and with which it is conceptually most related. The role operators, then, determine the order of the terms and thus provide the syntax of this language.

It is to be remembered that every string must contain at least one concept which is introduced by either the operator(1), representing ing, or else the operator(2) which represents an action. The preparation of the string is the ost important aspect of PRECIS indexing. For the actual display of entries, generation of epositions, etc., various manipulation codes are used in the preparation of the input. In this area PRECIS has developed some new techniques, the system of rotation of terms in the string for the preparation of the various entries and the logic behind such rotation is as follows :

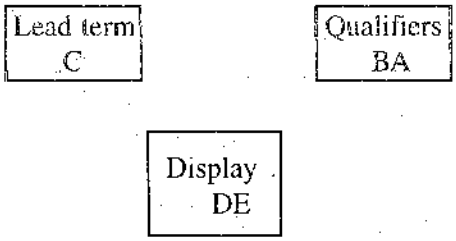
All the entries from the string are to be full entries unlike those in some systems where there is only one specific subject entry and all other entries are just reference entries only. Since all the component terms have to be retained in all the entries and the terms have to be rotated to bring them as lead terms, an additional device to preserve the context, as fixed in the string, is necessary. Suppose there are five terms in a string and they are syntactically related as :

A>B>C>D>E

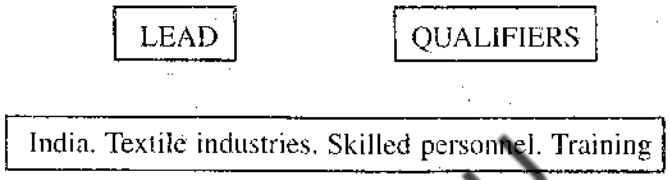
Now,if the terms C is the point of approach in a search formulation then the user of the index should be given the following view of the subject

<——————————>
A>B>C>D>E

This means that at C the user must know about its immediate context B and A, in that order, and the aspects of C that are being considered, that is, D and E, again in that order. This has been possible to achieve by introducing a simple two-line format for display. To meet the above search formulation the corresponding entry will be in the following format :



The method of rotation to generate the various entries can be compared to shunting. This can be demonstrated with an example. The example of the string that was prepared can be taken. Let us visualise that all the terms in the string are put in the display box in the correct order. This can be depicted as follows:



From the display box the terms are then to be pushed up, one by one like shunting. In the first shunting the term India will go up into the lead position and in the second shunting the term 'textile industries' will come to this position, sending back the previous term 'India' into the qualifier position. The following entries will demonstrate the shunting process :

India
Textile industries. Skilled personnel. Training

Textile industries. India
Skilled personnel. Training

Skilled personnel. Textile industries. India
Training

Training. Skilled personnel. Textile industries. India

It would be observed that all the four terms in the string have come, in turn, as lead term. The terms which are to be used as lead terms are to be decided by the indexer and this is indicated in the input by v. In the above example we have assumed that all the terms are to be used as lead terms.

It must be emphasised that all the four entries, given above, specify the same subject. The syntax is mainly controlled by the particular sequence of the terms and also the format for the entries. it has also been realised that sequence of terms alone may not bring out the correct meaning. To handle such situations, PRECIS has introduced a number of devices. Basically, the devices are the use of prepositions or prepositional phrases, slight variation in order of terms and display and introduction of facet-type concept term in the string.

7.2.4 Advantages and Limitations of PRECIS

ADVANTAGES :

1. PRECIS is the first type of subject indexing which uses a computer for the pre-coordinate indexing with a theory of chain set behind it.

2. The string format is set with numerals and alphabets (role operators) indicating the sequence.

3. It is an index system which generates cross references as well as subject thesaurus for different fields.

4. It solves the problem of avoiding generic entries and thereby reference to non-existing document.

5. For English as well as other languages PRECIS provides grammatically closet indexing technique.

6. It provides good associative links with Dewey Decimal Classification with a chain theory embedded in it.

LIMITATIONS :

From the user's point of view the following limitations are observed in PRECIS.

1. Use of incorrect prepositions or conjunctions by the users in between the pair of term at the time of search. Occasionally this has led to retrieval of non-relevant or partially relevant documents.

2. Difficulties are also found because of excessive use of the manipulation device. The rules of syntax need to be changed suitably to meet the specific problems of concept specification particularly with regard to members of quasi-generic group (role operator-q)

7.2.5 Examples of PRECIS

Subject : *Training of skilled personnel in textile industries of India*

Input String :

(0) France

(1) Textile industries

(p) Skilled personnel

(2) Training

Entries :

France
Textile industries.Skilled personnel.Training

Textile industries.France
Skilled personnel.Training

Skilled personnel. Textile industries.France
Training

Training. Skilled personnel.Textile industries.France

7.3 POSTULATE BASED PERMUTED SUBJECT INDEXING (POPSI)

7.3.1 Brief History and Development of POPSI

POPSI is a short form of Postulate-based Permuted Subject Indexing. The work on POPSI began in Documentation Research and Training Centre (DRTC), Bangalore in 1968. Thinking in this direction first began indirectly in 1964 and in 1966 an experiment was conducted for teaching purposes in DRTC. The result of this experiment being satisfactory, a research project was taken up in 1967. The project on this was completed in 1968. Since then continuous research on this new line of thinking (i.e. POPSI) is going on. Some of the results of this research have also been published. Some of the results of this research have also been published. Meanwhile POPSI has been used in practice also.

The postulates used in POPSI were those of Ranganathan. The application of these postulates to derive an alphabetical subject index was attempted in the form of classified index to collected works of Mahatma Gandhi and also to the Bibliography of Mahatma Gandhi. It was further developed by applying it to many subjects in social sciences.

From 1978 onwards, Bhattacharya worked towards the development of an indexing system by postulating a set of categories, namely, Discipline, Entity, Property, Action with Space and Time and modifiers. He also postulated the concepts of Core and Base for an index structure. The primary aim of Bhattacharya was to generate an indexing system that can generate in the verbal plane an organising classification on the basis of a coherent set of postulates about the semantic and syntactic structure of subject. His aim was also to generate on the basis of this organising classification an associate classification by using the technique of cyclic permutation of sought terms occurring in the modulated subject propositions.

Bhattacharya postulated a generalised subject indexing language with a set of elementary categories, namely, Discipline, Entity, Action, Property with a set of modifiers. Francis Devada son submitted his thesis on Computer based Systems for Generating different types of Subject Indexes and Alphabetical Classaurus based on the 'Deep Structure', of Subject Indexing Languages to Karnataka University in 1984. This Deep Structure Indexing System (DSIS) developed by Devadason is nothing but the computerised version of POPSI. Any study on POPSI should, therefore, be on DSIS as developed by Devadasan. The DSIS is based on the classification theory of Ranganathan, the great teacher in modern classification theory and

on the practical solution of chain indexing and the availability of a system like classaurus. With DSIS, Devadasan presents a solution for computerised string organisation in the world of string indexing in addition to the solution offered by Derek Austin's PRECIS and Tim Craven's different methods on computerised string organisation.

7.3.2 Features and Structure of POPSI

The DSIS is based on Deep Structure (DS) of Subject Indexing Language (SIL). It is based on

1) a set of postuplated Elementary Categories (EC of the elements fit to form components of names of subjects;

2) a set of syntax rules with reference to the categories;

3) a vocabulary control tool such as the Classaurus;

4) a set of indicator digits to denote the categories and their subdivisions; and

5) a set of codes to denote a few of the decisions of the indexer, in order to generate by computer manipulation, different types of subject/indexes

The DS of SIL postulates that the component ideas in the names of subjects can be deemed to fall into any one of the ECs Discipline (D), Entity (E), Property (P), and Action (A). In other words, a component idea can be a manifestation of only one of the ECs.

Subdivisions of Manifestation

Manifestations of each of the ECs may admit on subdivisions : Species/Type, Part and sometimes Constituent. A Species/Type does not disturb the conceptual wholeness of the manifestation to which it is Species/Type. A Part is a non-whole of the manifestation to which it is a part. A Constituent is an ultimate part with its own individuality. For exmmple, in the case of "House", /Multi-storeyed /House" is a Species/Type; 'Foundation', 'Roof', 'Door' 'Window' are parts; 'Cement', 'Mortar', 'Sand are Constituents. Constituents generally occupy for the EC Entity.

Modifier; Compound and Complex Terms: Apart from the ECs, a special component called Modifier is also recognised in the names of subjects. Modifier is an idea used or intended to be used to qualify (differentiate / speciate) the manifestation without disturbing its conceptual wholeness. For example, 'Red' in 'Red Rose'; 'Concrete in 'Concrete Bridge'. A Modifier generally creates Species/Type of the modifyee (focus). Modifiers can be Common Modifiers like Form, Time, Environment and Place/ and Special Modifiers, which can be based on /any of the ECs. Generally, Common Modifiers can modify a combination of two or more manifestations of two or more ECs.

Depending on the structure of the 'Modified Term', Modifiers could be further grouped into two types:

1) Modifier of Kind 1, that which requires auxiliary / function words to be inserted between the modifyee term and its modifier term forming a Complex Term. Eg: 'Finishing for Tip Effect', which is a type of 'Finishing' of Leather.

2) Modifier of Kind 2, that which does not require auxiliary /function words to be inserted in between, but automatically forms an acceptable Compound Term. Eg: 'Foot' forming the Compound Term 'Foot Bridge' which is a type of Bridge.

Note: The above grouping of modifiers depends on the natural language used for indexing. Moreover, that component in the name of subject represented as a Complex Term is likely to be changed into a Compound Term/term by subsequent emergence of new technical terms in the subject area concerned. The auxiliary/function words in complex terms may be role indicating words or 'phase relation' indicating words or prepositions, etc.

Composite Term

According to the postulate of ECs, any one component in the name of a subject can belong to any one (only one) of the ECs. If a component term represents a manifestations of more than one EC then it is a Composite (Category) Term. It should be broken down (factored, decomposed) into two or more constituent terms and each one of them should be identified as belonging to one or the other of the ECs.

The identification and factoring of Composite Terms is guided by the ECs of the indexing language. The Composite Term is considered in DSIS as a synonymous term to the combination of the factored constituent terms. For example,

Phthisis = Medicine (D) + Lung (E) + Tuberculosis (P)

Syntax of DS of SIL

The basic rule of syntax associated with the DS of SIL for formulating names of subjects is that, Discipline should be followed by Entity (both modified or unmodified) appropriately interpolated or extrapolated wherever warranted by Property and/or Action (both and/or modified or unmodified). In general, the rules of syntax give rise to the following sequence of components in a name of subject.

DISCIPLINE followed by ENTITY which is followed by PROPERTY and/or ACTION, PROPERTY and/or ACTION may be further followed by PROPERTY and/or ACTION as the case may be. Each of the above manifestations may further admit of, and be followed immediately by their respective SPECIES/TYPES and/or MODIFIERS and/or PARTS and/or CONSTITUENTS. The COMMON MODIFIERS generally occur last in the sequence.

The rules of syntax give rise to a context-dependent sequence of the components in the name of a subject in conformity with Ranganathan's Principles of Facet Sequence, such as the Actand - Action - Actor - Tool Principle.

Indicators of Deep Structure

Certain numeric codes have been prescribed in DSIS to indicate the manifestations of the different ECs, their subdivisions and modifiers of different kinds. These indicators are given below:

Common Modifiers

0 Form Modifier
2 Time Modifier
3 Environment Modifier
4 Place Modifier

Elementary Categories

9 Discipline
8 Entity
.2 Property
.1 Action

Subdivisions/Divisors

.3 Constituent
.4 Part
.5 Modifier of Kind 1 (including Phase Relation Modifier)
.6 Species/Type (including those created by Modifier of Kind 2)

In the name of a subject the indicators precede the components to which they are indicators. The indicators for Property and Action and also for the Subdivisions/Divisors are attached with the indicators for the ECs to which they are respectively Property or Action or Subdivisions/Divisors.

7.3.3 Application of POPSI

Taking the title as the starting point, names of each of the specific subjects dealt within the concerned document are expressed in natural language. Each of the component ideas corresponding to each of the ECs that are implied, are explicitly stated in each of the names of subjects to form an 'expressive title'. Let one of the expressive title be: In Leather Technology, *Evaluation of Two Bath Chrome Tanning of Leather Using Dichromate.*

Formalised Name of Subject

The expressive title is then analysed to identify the ECs, their subdivisions/divisors, to which each of the components in the expressive title belong. All Composite Terms are factored into their fundamental constituent terms and identified as belonging to one or the other of the ECs. The Composite Terms are noted separately for preparing Cross Reference (CR) entries to be included in the final index. The component terms are written down as a formalised expression following the rules of syntax, as given below:

"(Discipline) Leather Technology, (Entity) Leather, (Action on Entity) Two Bath Chrome Tanning, (Entity based Modifier of Kind1) (Using) Dichromate, (Action on Action) evaluation".

Modulated Name of Subject

Each of the component terms in the formalised name of subject is then analysed to find out its superordinate terms. This is done by finding out "of which the concerned component

is a Species/Type or Part of Constituent ?", in the context of the name of subject as a whole. This process is continued with each of such superordinates recognised in the process till it ends up with concept of the EC of which it is a manifestation. For this purpose terminological sources such as thesauri, classauri, dictionaries, etc are used. Each of the superordinates are fixed prior to the concerned term successfully giving rise to a `Modulated' name of subject as follows:

"(D) Leather Technology, (E) Leather, (A) Tanning, (Type of A)Mineral Tanning, (Type of A) Chrome Tanning, (Type of A) Two Bath Chrome Tanning (Modifier of Kind 1) (Using) Dichromate, (A on A) Evaluation".

Note: The reason for making each of the superordinates to precede the respective component terms is to endow the name of subject with the capacity to produce an organising sequence effect resembling the sequence of class numbers. Moreover, it is possible to prepare the alphabetical subject index using all or the relevant superordinate terms as Lead Terms, from the name of the subject itself. Generally, it is not necessary to 'modulate' Modifier of Kind 1 forming a complex term in the name of a subject.

Standardised Name of Subject

Each of the component terms in the name of a subject is replaced with standard terms and synonymous, quasi-synonymous terms are noted separately for preparing CR entries to be included in the index later. For this purpose vocabulary control tools such as thesauri, classauri, etc. are used.

Appropriate indicators for ECs, their subdivisions/divisors and Common Modifiers of different kinds are inserted in the appropriate places. The auxiliary/function words introducing Modifiers of Kind 1 are also standardised, if found necessary. The resulting name of subject is as follows:

"Leather Technology 8 Leather 8.1 Training 8.1.6 Mineral Tanning 8.1.6 Chrome Tanning 8.1.6 Two Bath Chrome Tanning 8.1.5 (using) Dichromate 8.1.1 Evaluation"

Note: The indicator digit for Discipline is not used as it is taken as understood to be the first digit in all names of subjects. In the component '8 Leather' the indicator '8' denotes that it is a manifestation of Entity. In the component '8.1 Tanning', the indicator '8.1' denotes that it is an Action on Entity, and so on. A set of 'Modulated' names of subjects with appropriate indicators when just sorted alphanumerically can produce an 'organising classification effect'. This has reduced considerably the See also CRs from narrower subjects/ terms to their respective broader subjects/terms (ascending references) and from broader subjects/terms to their respective narrower subjects/ terms (descending references).

Formation of Subject Headings

After formulating the names of a subject as per the DS of SIL and noting down the CR entries that are necessary, the indexer has to decide: 1) which component terms (including those forming a complex term) should form the Lead, and 2) which component terms should form the context, in order to produce headings for the different subject index entries.

Selection of Lead Terms

In order to provide access, significant terms in the name of a subject are selected to form the Lead Term by prefixing a process subject '$0' to the concerned term. Though it is difficult to ascertain which terms should form the Lead, certain guidelines could be followed. For instance, the term denoting the Discipline need not be selected to form the Lead, if the whole subject index is specifically for that Discipline alone. Moreover, very generic Entity terms such as Man, Plant, Animal, etc. very common Property terms such as Capacity, Efficiency, Cause, etc. very common Action terms such as Calculation, Determination, Evaluation, etc and terms denoting Common Modifiers need not necessarily be selected to form Lead Terms. But it should be decided by taking the name of the subject as a whole and the community to be served into consideration. It is helpful to select each of the Complex terms as such to form the Lead. A useful guideline for selecting Lead Terms is the Cannon of Sought Heading of Ranganathan.

Selection of Context Terms

The Context Heading sets the context in which the Lead Heading occurs. In order to provide the maximum context, the Context Heading in DSIS in general represents the full subject analysis along with the superordinates and indicators for each of the component terms. This is helpful in creating an organising sequence among the Context Headings to a particular Lead Heading.

If the purpose is only to serve the comprehending function then the superordinates included at the 'Modulation' step may be omitted, forming a 'Short Context Heading', provided it represents the full meaning of the subject. As it has been observed that "the syntactical role of each term in a heading is largely expressed by its position relative to the other terms and that in some cases, the position of a term is not of its own accord sufficient to indicate its role beyond a reasonable doubt, which means that an element of ambiguity will be present", the sequence of the terms in the Context Heading is kept invariant along with the different indicator digits in DSIS. For this purpose while selecting terms to form such 'Short Context Heading', it is necessary to select the last component term (it may be a Compound Term or a Complex Term) in each of the ECs and common modifiers. If the selected last component term does not by itself individualise it (which happens in general, when the selected term is a Part or a Constituent), then successive superordinate terms should also be selected so that it gets individualised and is homonym free. These superordinates are "Upper Links that resolve the homonym" and the process of their selection is similar to that of the Chain Indexing system. Terms selected to form Context Heading in DSIS are prefixed with a process code '$1'.

Upper Link Specifiers to Lead Term

A name of a subject formulated according to the DS of SIL can be considered as a Chain having as its links each of the component terms (Compound Terms and Complex Terms each taken as a unit component term). For a particular component term, all the other terms occurring prior (earlier) to it, when arranged according to the syntax rules, form upper links. When a term becomes the Lead Terms, some of the upper links could be suffixed to it to further specify the Lead Term. While selecting terms for forming Context Headings, care has been taken to see that the terms selected are such that the Short Context Heading is unambiguous and homonym free. Hence the upper links to a term under consideration, that are selected to

form Short Context Heading have been used to form Upper Link Specifiers to the concerned term when it becomes the Lead. The sequence of component terms in the Lead Heading containing Upper Link Specifiers taken from left to right is the reverse of the sequence of the terms arranged according to the rules of syntax. This 'reverse rendering' had been found necessary in retrieval.

Default Lead and Context

When a component term is neither selected to form Lead nor context, the default is that it is selected to form Lead. If a component term is selected to form only context, it is not considered for Lead at all. If a component term is the last component term of the EC manifestation, then it is selected to form the Context. This is done by checking the indicator digit with that of the succeeding term. If the difference in the indicator digits is only due to the succeeding term's indicator having a '6' denoting Species/Type, then no action is taken. To avoid a term (neither selected to form Lead nor Context) being considered for the default options, a null process code '$9' is prefixed to the concerned term. These default options are set automatically by computer manipulation.

Processing Codes

The following are the processing codes used in DSIS for computer manipulation: 1) '$0' - Lead Term' 2) '$1' - Context Term; 3) < (starter), '>' (arrester) - enclosed within, is a Complex Term; 4) '$2' - Lead in PCR arising out of Complex Term; 5) '$*' (auxiliary word identifier) '/' (auxiliary word delimiter) - enclosed within, is an auxiliary/function word(s); and 6) '$9' - neither Lead nor Context.

Apart from the above process codes, a special process code '$3' is used with Modifiers of Kind 2 to create Compound Terms automatically. Eg: "8 Skin 8.6$3 Pig 8.6$3 Cured" would produce "8 Skin 8.6Pig Skin 8.6 Cured Pig Skin".

Coding of the Name of Subject

The formulated name of a subject 'modulated' and 'standardised' given as example in section 3.3 is:

"Leather Technology 8 Leather 8.1 Tanning 8.1.6 Mineral Tanning 8.1.6 Chrome Tanning 8.1.6 Two Bath Chrome Tanning 8.1.5 (using) Dichromate 8.1.1 Evaluation".

In order to form the 'Short Context Heading', it is sufficient if the terms 'Leather', 'Two Bath Chrome Tanning (using) Dichromate' and 'Evaluation' are selected. As the term 'Leather' itself resolves any homonym that may arise, it is not necessary to select the Discipline term 'Leather Technology' to form the Short Context. These terms are prefixed with the process code '$1', indicating that they form (short) Context Heading. If we assume that the subject index is only for 'Leather Technology', then it is not necessary to select 'it' to form the Lead either. Hence, it is prefixed with the null process code '$9'. For the same reason, the term 'Leather' need not be selected to form Lead. However, the terms 'Tanning', 'Mineral Tanning', 'Chrome Tanning' and 'Two Bath Chrome Tanning (using) Dichromate' may be selected to form the Lead. These terms are prefixed with the process code '$0' to indicate that they are Lead Terms. The Complex Term is enclosed within angular brackets and the

auxiliary/ function word between '$*' and '/'. It may be necessary to form a PCR entry using the term 'Dichromate' and hence it is prefixed with the process code '$2'. The term 'Evaluation' is a very common Action term and it need not be selected to form the Lead. These decisions of the indexer are incorporated, to form the input name of a subject given below:

"$9 Leather Technology 8 $1 Leather 8.1 $0 Tanning 8.1.6 $0 Mineral Tanning 8.1.6 $0 Chrome Tanning 8.1.6 $0$1 < Two Bath Chrome Tanning 8.1.5 $* (using)/$2 Dichromate > 8.1.1$1 Evaluation".

The above input could be further simplified by taking default options whether applicable and using the process code '$3' for Modifier of Kind 2 to form Compound Term, as given below:

"$9 Leather Technology 8 $1 Leather 8.1 Tanning 8.1.6 $3 Mineral 8.1.6 Chrome Tanning 8.1.6 < Two Bath Tanning 8.1.5 $* (using)/$2 Dichromate > 8.1.1 $1 Evaluation".

Note: Most of the decisions relating to Lead and Context terms selection could be left to the default options available. But synonyms, quasi-synonyms and synonyms due to 'factoring' of Composite Terms are to be noted separately to for CR entries to be included in the index before final sorting and printing.

Entries:

Uni-Component Term Lead Heading with Full Context Heading

TANNING

LEATHER TECHNOLOGY 8.LEATHER 8.1 TANNING 8.1.6 MINERAL TANNING 8.1.6 CHROME TANNING 8.1.6 TWO BATH CHROME TANNING 8.1.5 (USING) DICHROMATE 8.1.1 EVALUATION

MINERAL TANNING

LEATHER TECHNOLOGY 8.LEATHER 8.1.TANNING 8.1.6 MINERAL TANNING 8.1.6 CHROME TANNING 8.1.6 TWO BATH CHROME TANNING 8.1.5 (USING) DICHROMATE 8.1.1 EVALUATION

CHROME TANNING

LEATHER TECHNOLOGY 8.LEATHER 8.1 TANNING 8.1.6 MINERAL TANNING 8.1.6 CHROME TANNING 8.1.6 TWO BATH CHROME TANNING 8.1.5 (USING) DICHROMATE 8.1.1 EVALUATION

TWO BATH CHROME TANNING (USING) DICHROMATE

LEATHER TECHNOLOGY 8.LEATHER 8.1 TANNING 8.1.6 MINERAL TANNING 8.1.6 CHROME TANNING 8.1.6 TWO BATH CHROME TANNING 8.1.5 (USING) DICHROMATE 8.1.1 EVALUATION

7.3.4 Advantages and Limitations of POPSI

Advantages:

1) It provides a flexible approach to subject indexing

2) It is amenable to any variety of transformations

3) The computer aided DSIS is quite simple because the indexer's decisions are quite few and the process codes used are also few

4) The system provides the facility of lending some of the indexer's decisions to default options.

5) It provides for the generation of different types of subject index entries and also the vocabulary control tool classaurus using computer.

6) It helps to keep up-to-date the vocabulary control tool classaurus used for standardising the name of subject and for preparing the necessary cross reference entries.

7) DSIS has a programme to create a bibliographic database capable of handling twelve variable length data fields using a simple directory of pointers at the beginning of each record.

Limitations

The following disadvantages are found by the users of POPSI system:

1) Presence of Discipline concept as the first element in the context leads to the possibility of using this Discipline concept as a coordinate concept and thus it becomes a redundant concept at the time of search.

2) Use of numerals in index entries reduces the clarity of index entries

3) The similarity of lead and context headings becomes unnecessary on many occasions in index entries.

7.3.5 Examples of POPSI

Subject: Evaluation of two bath chrome tanning of leather using Dichromate

Formalised Name of Subject

"(Discipline) Leather Technology, (Entity) Leather, (Action on Entity) Two Bath Chrome Tanning, (Entity based Modifier of Kind 1) (Using) Dichromate, (Action on Action) Evaluation"

...er, (A) Tanning, (Type of A) Mineral Tanning,

(Type of A) Chrome Tanning, (Type of A) Two Ba
(Using) Dichromate, (A on A) Evaluation".

Standardised Name of Subject

"Leather Technology 8 Leather 8.1 Tanning 8.
8.1.6 Two Bath Chrome Tanning 8.1.5 (using) Dich

Entries:

Uni-Component Term Lead Heading with Full

TANNING

LEATHER TECHNOLOGY 8.LEATH
TANNING 8.1.6 CHROME TANNING
8.1.5 (USING) DICHROMATE 8.1.1 EV

MINERAL TANNING

LEATHER TECHNOLOGY 8.LEATH
TANNING 8.1.6 CHROME TANNING
8.1.5 (USING) DICHROMATE 8.1.1 EV

CHROME TANNING

LEATHER TECHNOLOGY 8.LEATHE
TANNING 8.1.6 CHROME TANNING 8
8.1.5 (USING) DICHROMATE 8.1.1 EV

TWO BATH CHROME TANNING (USING) DICH

LEATHER TECHNOLOGY 8.LEATHE
TANNING 8.1.6 CHROME TANNING 8
8.1.5 (USING) DICHROMATE 8.1.1 EV

7.4 COMPARATIVE STUDY OF

PRECIS and POPSI are both rotated indexes.
permutation and devices to preserve the context
preservation of context in both the systems is di
preserved in the following manner.

If the concepts A, B, C, and D in a string are
C is the approach terms, ie., lead them the context

C>B>A

In case of POPSI the context is preserved in the following way: A>B>C>D

1) It is observed that the syntax of PRECIS index entries is more logical than DSIS

2) Both the systems are applicable with equal efficiency to the literates of social, pure and applied sciences.

7.5 LET US SUM UP

Indexing is an essential process to help information retrieval in manual as well as computer-based systems. In the world of indexing PRECIS developed by Derek Austin of BNB and POPSI, the brain-child of Dr S.R. Ranganathan and further developed by Prof. G. Bhattacharya and computerised by Dr F.J. Devadason of DRTC have commanded high esteem. Both PRECIS and POPSI are having equal efficiency in their expressiveness. In the matter of syntax the PRECIS syntax is more logical than POPSI. But both the systems can very well be applied to the literature of pure and applied sciences and social sciences.

7.6 REFERENCES

1) BHATTACHARYYA, G. "Fundamentals of subject indexing languages". IN *Proceedings of the third international study conference on classification research* (Jan 6-11, 1975:Bombay). Bangalore: DRTC, 1979. p.86-98.

2) BHATTACHARYYA, G. "Some significant results of current classification research in India". IN *International Forum for Information and Documentation* 6(1981); No.1; p.11

3) SPANG-HANSSEN, H. "Are classification systems similar to natural languages ?" IN *Proceedings of the third international study conference on classification research* (Jan 6-11, 1975:Bombay). Bangalore: DRTC, 1979. p.15

4) GARDIN, J.C. "Document analysis and linguistic theory". *Journal of Documentation* 29 (2); 1973. p.146-7

5) BHATTACHARYYA, G. "Foreword to Fugmann, R: The analytico-synthetic foundation for large indexing and information retrieval systems". IN *Sarada Ranganathan Endowment for Library Science.* 1983. p.IX

6) BHATTACHARYYA, G. "POPSI: Its fundamentals and procedure based on a general theory of subject indexing languages". *Library Science Slant to Documentation* 16 (1); 1979. p.14-15

7) BHATTACHARYYA, G. A *general theory of subject indexing language* (Ph.D. thesis). Dharward: Karnataka University, 1980.

8) RANGANATHAN, S.R. *Prolegomena to library classification.* 3rd ed. Bombay: Asia Publishing House, 1967. p.422-424

9) BHATTACHARYYA, G. "POPSI: a source language for organising and associative classification". *Library Science Slant to Documentation* 19; 1982. p.249-252

10) KAISER, J. "Systematic indexing". *Aslib Report of Proceedings.* Oxford: Sept. 24-27, 1926. p.20-33. [Reprinted: Readings in library cataloguing/ edited by R.K. Olding. Canberra, AU: F.W. Cheshire Ltd, 1966. p.154; Reprinted again: New Delhi: Lakshmi Book House, 1967. p.154]

11) RANGANATHAN, S.R.*Prolegomena to library classification.* London: Edward Goldston Ltd, 1944. p.39

12) RANGANATHAN, S.R. *Elements of library classification.* Bombay: Asia, 1962. p.130

13) AUSTIN, Derek and Jeremy A Digger. "PRECIS: The Preserved Context Index System". *Library Resources and Technical Services.* 21:13-30; Winter 1977.

AUSTIN, Derek and Jutta Sorensen. *PRECIS Training Course Material,* 1978.

7.7 RECOMMENDED BOOKS

AUSTIN, Derek. *PRECIS: A Manual of Concept Analysis and Subject Indexing.* London: Council of the British National Bibliography, 1984.

FOSKET, A.C. *Subject Approach to Information.* 4th ed. London: Clive Bingley, 1982.

*INDEXING systems: Concepts, models and techniques/*edited by T.N. Rajan. Calcutta: IASLIC, 1981.

PRASHER, R.G. *Index and Indexing Systems.* New Delhi: Medalion Press, 1989.

RANGANATHAN, S.R. *Prolegomena to library classification,* 3rd ed. Bombay: Asia Publishing House, 1937.

RANGANATHAN, S.R. *Classified Catalogue Code with Additional Rules for Dictionary Catalogue Code.* Bombay: Asia Publishing House, 1964.

VARMA, A.K. *Trends in Subject Indexing.* Delhi: Mittal Publications, 1984.

7.8 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) Describe the syntactical relationship in PRECIS and POPSI
2) Make a comparative study of PRECIS and POPSI. Give suitable examples.
3) Discuss the salient features of PRECIS
4) Make out an analysis of the syntactic structure of DSIS
5) Discuss the salient features of POPSI

II SHORT NOTES

a) Syntax in PRECIS

b) Subject Indexing

UNIT - 8 : THESAURUS - ITS STRUCTURE, FUNCTIONS AND CONSTRUCTION

Structure

8.0 AIMS AND OBJECTIVES

Thesaurus, as a vocabulary control device, has been identified as an important tool in information processing and retrieval. The present unit aims to provide an overview of its structure, functions and construction.

auxiliary/ function word between '$*' and '/'. It may be necessary to form a PCR entry using the term 'Dichromate' and hence it is prefixed with the process code '$2'. The term 'Evaluation' is a very common Action term and it need not be selected to form the Lead. These decisions of the indexer are incorporated, to form the input name of a subject given below:

"$9 Leather Technology 8 $1 Leather 8.1 $0 Tanning 8.1.6 $0 Mineral Tanning 8.1.6 $0 Chrome Tanning 8.1.6 $0$1 < Two Bath Chrome Tanning 8.1.5 $* (using)/$2 Dichromate > 8.1.1$1 Evaluation".

The above input could be further simplified by taking default options whether applicable and using the process code '$3' for Modifier of Kind 2 to form Compound Term, as given below:

"$9 Leather Technology 8 $1 Leather 8.1 Tanning 8.1.6 $3 Mineral 8.1.6 Chrome Tanning 8.1.6 < Two Bath Tanning 8.1.5 $* (using)/$2 Dichromate > 8.1.1 $1 Evaluation".

Note: Most of the decisions relating to Lead and Context terms selection could be left to the default options available. But synonyms, quasi-synonyms and synonyms due to 'factoring' of Composite Terms are to be noted separately to for CR entries to be included in the index before final sorting and printing.

Entries:

Uni-Component Term Lead Heading with Full Context Heading

TANNING

LEATHER TECHNOLOGY 8.LEATHER 8.1 TANNING 8.1.6 MINERAL TANNING 8.1.6 CHROME TANNING 8.1.6 TWO BATH CHROME TANNING 8.1.5 (USING) DICHROMATE 8.1.1 EVALUATION

MINERAL TANNING

LEATHER TECHNOLOGY 8.LEATHER 8.1.TANNING 8.1.6 MINERAL TANNING 8.1.6 CHROME TANNING 8.1.6 TWO BATH CHROME TANNING 8.1.5 (USING) DICHROMATE 8.1.1 EVALUATION

CHROME TANNING

LEATHER TECHNOLOGY 8.LEATHER 8.1 TANNING 8.1.6 MINERAL TANNING 8.1.6 CHROME TANNING 8.1.6 TWO BATH CHROME TANNING 8.1.5 (USING) DICHROMATE 8.1.1 EVALUATION

TWO BATH CHROME TANNING (USING) DICHROMATE

LEATHER TECHNOLOGY 8.LEATHER 8.1 TANNING 8.1.6 MINERAL TANNING 8.1.6 CHROME TANNING 8.1.6 TWO BATH CHROME TANNING 8.1.5 (USING) DICHROMATE 8.1.1 EVALUATION

7.3.4 Advantages and Limitations of POPSI

Advantages:

1) It provides a flexible approach to subject indexing

2) It is amenable to any variety of transformations

3) The computer aided DSIS is quite simple because the indexer's decisions are quite few and the process codes used are also few

4) The system provides the facility of lending some of the indexer's decisions to default options.

5) It provides for the generation of different types of subject index entries and also the vocabulary control tool classaurus using computer.

6) It helps to keep up-to-date the vocabulary control tool classaurus used for standardising the name of subject and for preparing the necessary cross reference entries.

7) DSIS has a programme to create a bibliographic database capable of handling twelve variable length data fields using a simple directory of pointers at the beginning of each record.

Limitations

The following disadvantages are found by the users of POPSI system:

1) Presence of Discipline concept as the first element in the context leads to the possibility of using this Discipline concept as a coordinate concept and thus it becomes a redundant concept at the time of search.

2) Use of numerals in index entries reduces the clarity of index entries

3) The similarity of lead and context headings becomes unnecessary on many occasions in index entries.

7.3.5 Examples of POPSI

Subject: Evaluation of two bath chrome tanning of leather using Dichromate

Formalised Name of Subject

"(Discipline) Leather Technology, (Entity) Leather, (Action on Entity) Two Bath Chrome Tanning, (Entity based Modifier of Kind 1) (Using) Dichromate, (Action on Action) Evaluation"

er, (A) Tanning, (Type of A) Mineral Tanning,

In case of POPSI the context is preserved in the following way: A>B>C>D

1) It is observed that the syntax of PRECIS index entries is more logical than DSIS

2) Both the systems are applicable with equal efficiency to the literates of social, pure and applied sciences.

7.5 LET US SUM UP

Indexing is an essential process to help information retrieval in manual as well as computer-based systems. In the world of indexing PRECIS developed by Derek Austin of BNB and POPSI, the brain-child of Dr S.R. Ranganathan and further developed by Prof. G. Bhattacharya and computerised by Dr F.J. Devadason of DRTC have commanded high esteem. Both PRECIS and POPSI are having equal efficiency in their expressiveness. In the matter of syntax the PRECIS syntax is more logical than POPSI. But both the systems can very well be applied to the literature of pure and applied sciences and social sciences.

7.6 REFERENCES

1) BHATTACHARYYA, G. "Fundamentals of subject indexing languages". IN *Proceedings of the third international study conference on classification research* (Jan 6-11, 1975:Bombay). Bangalore: DRTC, 1979. p.86-98.

2) BHATTACHARYYA, G. "Some significant results of current classification research in India". IN *International Forum for Information and Documentation* 6(1981); No.1; p.11

3) SPANG-HANSSEN, H. "Are classification systems similar to natural languages ?" IN *Proceedings of the third international study conference on classification research* (Jan 6-11, 1975:Bombay). Bangalore: DRTC, 1979. p.15

4) GARDIN, J.C. "Document analysis and linguistic theory". *Journal of Documentation* 29 (2); 1973. p.146-7

5) BHATTACHARYYA, G. "Foreword to Fugmann, R: The analytico-synthetic foundation for large indexing and information retrieval systems". IN *Sarada Ranganathan Endowment for Library Science.* 1983. p.IX

6) BHATTACHARYYA, G. "POPSI: Its fundamentals and procedure based on a general theory of subject indexing languages". *Library Science Slant to Documentation* 16 (1); 1979. p.14-15

7) BHATTACHARYYA, G. A *general theory of subject indexing language* (Ph.D. thesis). Dharward: Karnataka University, 1980.

8) RANGANATHAN, S.R. *Prolegomena to library classification.* 3rd ed. Bombay: Asia Publishing House, 1967. p.422-424

9) BHATTACHARYYA, G. "POPSI: a source language for organising and associative classification". *Library Science Slant to Documentation* 19; 1982. p.249-252

(Type of A) Chrome Tanning, (Type of A) Two Bath Chrome Tanning (Modifier of Kind 1) (Using) Dichromate, (A on A) Evaluation".

Standardised Name of Subject

"Leather Technology 8 Leather 8.1 Tanning 8.1.6 Mineral Tanning 8.1.6 Chrome Tanning 8.1.6 Two Bath Chrome Tanning 8.1.5 (using) Dichromate 8.1.1 Evaluation".

Entries:

Uni-Component Term Lead Heading with Full Context Heading

TANNING

LEATHER TECHNOLOGY 8.LEATHER 8.1 TANNING 8.1.6 MINERAL TANNING 8.1.6 CHROME TANNING 8.1.6 TWO BATH CHROME TANNING 8.1.5 (USING) DICHROMATE 8.1.1 EVALUATION

MINERAL TANNING

LEATHER TECHNOLOGY 8.LEATHER 8.1 TANNING 8.1.6 MINERAL TANNING 8.1.6 CHROME TANNING 8.1.6 TWO BATH CHROME TANNING 8.1.5 (USING) DICHROMATE 8.1.1 EVALUATION

CHROME TANNING

LEATHER TECHNOLOGY 8.LEATHER 8.1 TANNING 8.1.6 MINERAL TANNING 8.1.6 CHROME TANNING 8.1.6 TWO BATH CHROME TANNING 8.1.5 (USING) DICHROMATE 8.1.1 EVALUATION

TWO BATH CHROME TANNING (USING) DICHROMATE

LEATHER TECHNOLOGY 8.LEATHER 8.1 TANNING 8.1.6 MINERAL TANNING 8.1.6 CHROME TANNING 8.1.6 TWO BATH CHROME TANNING 8.1.5 (USING) DICHROMATE 8.1.1 EVALUATION

7.4 COMPARATIVE STUDY OF PRECIS AND POPSI

PRECIS and POPSI are both rotated indexes. Both of them use methods of rotation or permutation and devices to preserve the contexts of the terms in the heading. But the preservation of context in both the systems is different. In case of PRECIS the context is preserved in the following manner.

If the concepts A, B, C, and D in a string are related in the manner A>B>C>D and if C is the approach terms, ie., lead them the context is preserved in the following manner:

C>B>A

UNIT - 8 : THESAURUS - ITS STRUCTURE, FUNCTIONS AND CONSTRUCTION

Structure

8.0 AIMS AND OBJECTIVES

Thesaurus, as a vocabulary control device, has been identified as an important tool in information processing and retrieval. The present unit aims to provide an overview of its structure, functions and construction.

10) KAISER, J. "Systematic indexing". *Aslib Report of Proceedings.* Oxford: Sept. 24-27, 1926. p.20-33. [Reprinted: Readings in library cataloguing/ edited by R.K. Olding. Canberra, AU: F.W. Cheshire Ltd, 1966. p.154; Reprinted again: New Delhi: Lakshmi Book House, 1967. p.154]

11) RANGANATHAN, S.R.*Prolegomena to library classification.* London: Edward Goldston Ltd, 1944. p.39

12) RANGANATHAN, S.R. *Elements of library classification.* Bombay: Asia, 1962. p.130

13) AUSTIN, Derek and Jeremy A Digger. "PRECIS: The Preserved Context Index System". *Library Resources and Technical Services.* 21:13-30; Winter 1977.

AUSTIN, Derek and Jutta Sorensen. *PRECIS Training Course Material,* 1978.

7.7 RECOMMENDED BOOKS

AUSTIN, Derek. *PRECIS: A Manual of Concept Analysis and Subject Indexing.* London: Council of the British National Bibliography, 1984.

FOSKET, A.C. *Subject Approach to Information.* 4th ed. London: Clive Bingley, 1982.

*INDEXING systems: Concepts, models and techniques/*edited by T.N. Rajan. Calcutta: IASLIC, 1981.

PRASHER, R.G. *Index and Indexing Systems.* New Delhi: Medalion Press, 1989.

RANGANATHAN, S.R. *Prolegomena to library classification,* 3rd ed. Bombay: Asia Publishing House, 1937.

RANGANATHAN, S.R. *Classified Catalogue Code with Additional Rules for Dictionary Catalogue Code.* Bombay: Asia Publishing House, 1964.

VARMA, A.K. *Trends in Subject Indexing.* Delhi: Mittal Publications, 1984.

7.8 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) Describe the syntactical relationship in PRECIS and POPSI
2) Make a comparative study of PRECIS and POPSI. Give suitable examples.
3) Discuss the salient features of PRECIS
4) Make out an analysis of the syntactic structure of DSIS
5) Discuss the salient features of POPSI

II SHORT NOTES

a) Syntax in PRECIS

b) Subject Indexing

After studying this unit, you will be in a position to

- explain the need, purpose and functions of a thesaurus
- describe the structure of a thesaurus
- explain the construction of a thesaurus through manual methods as well as computerised systems
- discuss the role of thesaurus in information storage and retrieval systems.

8.1 INTRODUCTION

The word 'Thesaurus' is derived from Greek word 'thesaurus', which mean a treasury. According to the *Oxford English Dictionary*, the earliest usage of word thesaurus was known in 1565 from the title, *Thesaurus Linguae Romanae et Britannicae*. While its first English usage was given in 1736. In 1852, Peter Mark Roget published his *Thesaurus of English Words and Phrases*, with the subtitle, *a collection of words classified and arranged so as to facilitate the expression of ideas and to assist in literary composition*. It is considered as the "Father of all Thesauri", though the concept has passed through many successive stages of metamorphosis over the past two centuries. Spark-Jones has studied deeply and distinguished various intertwined strands in the thesauri history. She traces the origin of synonymy in the subject classification of vocabularies, viz., *Amara Cosha*, etc.

The concept of Thesaurus entered the library and information science field in the early 1950s. Helen Brownson used the word for the first time in connection with information retrieval at the Dorking Conference on Classification on May 14, 1957. Though H.P. Luhn's *'Dictionary of Notions and Notational Familities'* and the *'Word-association Matrix'* of all were the earlier attempts, even before the conference. The 1960s has witnessed the development of major thesauri for the information retrieval. Sooner, the multilingual thesauri were also developed based on the Roget's *Polyglot Lexicon*.

8.2 THESAURUS - DEFINITIONS AND MEANING

Before discussing the structure and functions of a thesaurus, let us first examine some of the important definitions of the term 'Thesaurus' and understand the meaning of them.

There are several definitions that are prevalent in the reference tools for the term Thesaurus. Though it is not intended to provide a comprehensive list of definitions available in the published reference tools, however, some of the important definitions are quoted and discussed with a view to bring out a common thread that runs through all these definitions.

The Oxford English Dictionary defines the Thesaurus as a 'treasury' or 'storehouse' of knowledge as a dictionary, encyclopedia, or the like.

According to *Webster's Third New International Dictionary of English Language*, thesaurus

means, "a book containing a store of words of information about a particular field or set of concepts, specifically, a dictionary of synonyms".

According to the definitions used in that of the World Science Information System of UNESCO (namely, UNISIST), a thesaurus may be defined either in terms of its function or its structure.

In terms of function, a thesaurus is a terminological control device used in translating from the natural language of documents, 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.

The second revised edition of the *Guidelines for the Establishment and Development of Monolingual Thesauri,* defines the thesaurus as "the vocabulary of a controlled indexing language, formerly organised so that the *a priori* relationships between the concepts (e.g., as 'broader' and 'narrower') are made explicit".

All the above definitions show that the thesaurus means a treasury or storehouse in its general usage. As the word entered the literary field, through the wellknown Roget's Thesaurus, the usage has been changed to mean a lexicon or a dictionary or encyclopedia, where a set of concepts or words or phrases of a particular field are grouped together according to similarities in their meaning. The 'words with similar meaning' implies the synonymy, hence these thesauri are also called 'Synonymous Dictionaries'. Such dictionaries became reference tools in a literary field.

The usage of the term, Thesaurus, in LIS reference tools denotes a special function in information retrieval through controlling the terminology. Hence, the structure has also been changed to display relationships within the vocabulary based on semantics, not on orthography. This is an advanced feature over the mere grouping of synonyms. Nevertheless, the definition given in the first edition of the *UNISIST Guidelines for the Establishment and Development of Monolingual Thesauri* faced a great deal of criticism over the terms 'system language', 'documentation language' and 'information language' and hence, the second rerevised edition brought into use the proper name 'Indexing Language'.

8.3 STRUCTURE OF A THESAURUS ENTRY

A thesaurus should include the terms which represent the various concepts of a subject. The terms, which represent the concepts are called Descriptors. The descriptors are arranged in thesaurus an alphabetical order. An indexer assigns the descriptor terms to describe the contents of documents. The terms which are not preferred to be used in indexing are 'Non-descriptors'. They are proper names of corporate bodies, government agencies, institutions and firms, geographical names, etc. in addition, scope notes and definitions are also given.

Three kinds of interrelationships between the concepts are usually displayed in a thesaurus. They are - Hierarchical, Equivalence and Associative Relationships. The latter two types may be grouped under Non-Hierarchical Relationship.

a) ***Hierarchical or Structural Relationship:*** The hierarchical relation expresses super/ subordinate of concepts. There are two types of relationships in this category: Genus-Species and Part-Whole relationships. Relation between genus and its species and the part-whole relationships are displayed in the thesaurus by using the symbols BT (Broader Term) and NT (Narrower Term); and

b) ***Equivalence or Preferential Relations:*** When terms are regarded as similar or almost the same in meaning, they can be combined with the same concept. Synonyms are indicated in the thesaurus by the terms USE and UF (Used For).

c) ***Associative or Affinitive Relation:*** This relationship is employed to cover other relationships between concepts that are related but are neither consistently hierarchical nor equivalent. This relationship in a thesaurus is displayed by the symbol RT (Related Term).

A typical entry taken from *Thesaurus of Engineering and Scientific Terms* (TEST), indicating all the three relationships is shown below:

POLICE
UF Bodyguards (Personnel)

BT PERSONNEL
NT MILITARY POLICE
RT INDUSTRIAL PLANT PROTECTION
INTERNAL SECURITY

Thus, a thesaurus contains terms so arranged as to express their structural and functional relationships. A list of terms which does not include structural and relational information is not a thesaurus. It is merely an alphabetical list of descriptors or subject headings.

Generally, a thesaurus includes two parts. They are main part and an auxiliary part.

The ***Main Part*** in a thesaurus is a normal alphabetical list of all descriptors giving complete information on each descriptor, including the concept relationship. This part includes both descriptors and non-descriptors along with scope notes and definitions.

In order to improve the access to the main part, a thesaurus may contain several ***auxiliary parts,*** i.e., Permuterm Subject Index, Systematic Listings like Hierarchical Index, and Subject Category Index as in Thesaurus of Engineering and Scientific Terms (TEST) and graphic display of relationships as in SPINES Thesaurus. The thesaurus may also contain a faceted classification along with alphabetical thesaurus as in Thesaurofacet.

8.4 TYPES AND FUNCTIONS OF THESAURUS

The use of thesaurus in modern information storage and retrieval systems has been increasing. We need to know the types of thesauri and understand the functions of a thesauri.

8.4.1 Types of Thesaurus

Thesauri can be categoried into different types based on different characteristics they possess.

I Microthesaurus Vs. Macrothesaurus

Based on the scope of the subject field, thesauri can be divided into two types: Microthesaurus and Macrothesaurus. Macrothesaurus covers relatively a broad subject like science, engineering, social sciences, or even the interdisciplinary subjects such as development. The Microthesaurus on the other hand includes a smaller subjects like machine tools, toy marketing, etc.

II Structured vs. Unstructured

Based on the syntax of the indexing language, the thesauri can be categorised as i) Unstructured Thesauri, and ii) Structured Thesauri.

The Unstructured thesauri contain the unstructured vocabulary of uniterms and uniconcepts. This may give good recall, but is likely to result in low relevance. The often quoted example is 'Venetian Blind', when factored could lead to errors, which may be described as 'False Drops'.

The structured thesauri is associated with a classification scheme and functions as alphabetical index. It can be used independently of the classification scheme. The terms are treated as single semantic units and hence some sort of pre-coordination is involved.

III Using Preferred Terms Vs. Not Using Preferred Terms

Another way of categorising the thesauri is on the basis of nature of terminological control they adopt. These are two types: i) Thesauri using Preferred Terms, and ii) Thesauri not using Preferred Terms. In the first one, only one term denoting a concept is permitted for indexing and retrieval. These thesauri perform terminological control by preferred terms. In the second type, all the terms denoting a concept are allowed to be used for indexing and retrieval. When a term changes its meaning, it cancels the earlier synonyms and concept relationships. It requires computer systems for maintenance and retrieval.

IV Source, Adjunct and Cumulative Thesauri

Thesauri can also be categorised into three types by their nature of construction. i) Source Thesaurus, ii) Adjunct Thesaurus, and iii) Cumulative Thesaurus.

Source Thesaurus: As the name indicates, it acts as source or a databank from which an indexing tool can be extracted. Besides terminology, it also contains a guidance structure to terminology and classification, which will be helpful for the cooperating institutions to adopt it for constructing a thesaurus or indexing tool for specific application. Hence, Atchison and Gilchrist called it a 'Convertible Thesaurus'.

Adjunct Thesaurus: An adjunct thesaurus deals with a specific facet that does not show many relationships between concepts. It cannot be used independently and has to be added to a main thesaurus, hence it is called as an Adjunct Thesaurus.

Cumulative Thesaurus: A Cumulative Thesaurus collects information contained in a different thesauri or classification schemes and cumulates the same in its construction.

V Thesaurofacet

The thesaurofacet is a combined approach of a faceted classification and a thesaurus. A thesaurofacet was published in 1970 by Jean Aitchison under the title, *Thesaurofacet: a thesaurus and faceted classification for Engineering and related subjects.* The EE Classification for Engineering is a faceted classification and uses the synthesised notation. The notation is mixed and non-expressive. The classification scheme can only display one set of genus-species divisions, while BT-NT-RT cross references are shown in thesaurus. The whole system depends on the interaction of classification schedule and thesaurus.

8.4.2 Functions of a Thesaurus

The Encyclopedia of Library and Information Science (Vol.30) has enumerated the major purposes of a thesaurus, which can be summarised as follows:

1) To provide a map of given field of knowledge, indicating how concepts or ideas about concepts are related to one another, which helps an indexer or a searcher to understand the structure of the field;

2) To provide a standards vocabulary for a given subject field which will ensure that indexers are consistent when they are making index entries to an information storage and retrieval system;

3) To provide a system of references between terms which will ensure that only one term from a set of syhonyms is used for indexing concept;

4) To provide a guide for users of the system so that they choose the correct term for a subject search;

5) To locate new concepts in a scheme of a relationships with existing concepts in a way which makes sense to users of the system;

6) To provide classified hierarchies so that a search can be broadened or narrowed systematically.

Thesaurus is only a tool used for indexing. The tool has to be used only to control the vocabulary but not for subject analysis. The tool plays the following two roles: i) Prescriptive role: It prescribes as to what term should be assigned, and ii) Suggestive role: It suggests terms to be considered instead of, or in addition to, etc. The RT and to a certain extent BT and NT references are suggestive indicators.

A thesaurus helps in indexing the collections of documents, abstracts, bulletins, bibliographic and current awareness tools, etc. The standardised terminology helps to provide consistent representation of subject matter as index terms (in input) and also in searching and controlling the output. It also facilitates the manipulation of searches either broadening or narrowing it (generic search). It also widens the scope of the searches by bringing together the terms that are semantically related.

8.4.3 Role of Thesaurus in ISAR

Vickery (1965) has identified three stages of operation in the process of indexing. They are i) The text is scanned to select a set of words, phrases or sentences which collectively represent its subject, ii) A decision is taken as to which of these subject descriptors are worth recording as being relevant to the purpose of the retrieval system, and iii) The subject descriptions are transferred into the standard descriptor language used in the system. There has been a hunt for a standard descriptor language which can act as the main agent in establishing a coincidence of vocabularies in indexing and searching by conducting both operations in a common language.

The limitations of the classification and indexing systems are more prominent in their failure to identify the semantic and syntactic relationships of conceptual terms. Thesaurus has become more prominent as it exploits to the maximum and stands distinguished from the other categories of tools such as subject dictionaries/glossaries, alphabetic subject headings, classification schemes, etc.

Thesauri has been recognised and used as an effective indexing technique for more than past three decades to improve the precision and recall in information retrieval systems. In an environment of growing inter-disciplinarity of the subject fields as well as the automated information storage, processing and retrieval systems, the need for thesauri has been growing.

Helen Brownson said, "the problem (of information retrieval) ... is to transform concepts and relationships, as expressed in the language of documents, into a somewhat more regularised language, with synonyms controlled and syntactic structures simplified. Some investigators have come up with the thought that the best answer... may be the application of mechanised thesauri based on networks of related meanings".

Therefore, many integrated databases even at the international level use their authority/ descriptor lists in their subject fields besides free text searching techniques.

8.5 CONSTRUCTION OF THESAURUS

In the 1970s and 80s, several library and information scientists like Aitchison and Gilchrist (1972), Tonley and Gee (1980) and many others have suggested methods for construction of a thesaurus through manual methods. The PGI/UNISIST (1981) prepared guidelines for the construction of monolingual thesauri. Since the early 1980s, semi-automatic or computerised construction of thesauri became popular.

8.5.1 Steps in the Construction of Thesaurus

Aitchison and Gilchrist have given some practical steps for construction of a thesaurus.

They are further elaborated as below:

1) Identify the needs of the users
2) Define the subject field
3) Decide the type and design of thesaurus layout
4) Collect terms from subject literature, users, specialists
5) Screen and edit the terms as per the rules of thesaurus
6) Record the terms in the Terms Cards / Thesaurus Form
7) Sorting and Grouping of Thesaurus Cards
8) Prepare the hierarchical structure and other associated parts
9) Test the thesaurus against a selected collection of documents
10) Get the thesaurus evaluated by subject specialists and users

Let us further discuss these ten steps in detail:

1) Identification of the User Needs: A basic step for realising the need for design and construction a thesaurus, presupposes the needs of the users of a library and information system. Lack of proper indexing system or inadequacy of the existing retrieval systems may prompt the need for designing a thesaurus.

2) Definition of the Subject Field: Having decided the need for constructing a thesaurus, we have to define the subject field of the thesaurus, by establishing the boundaries of the subject. It should include the core/central area as well as the marginal/peripheral subject areas. Study of the general or technical thesauri already available may give some idea about the main sections of the subject field. The sections which are relevant, especially for the peripheral subject areas, could be adopted. Form and geographical divisions could be taken from any major classification systems or thesauri.

3) Decide the Type and Design of the Thesaurus Layout : At this stage, the compiler should clarify his ideas on the type of thesaurus to be constructed. Decision should be taken on the characteristics of thesaurus, such as specificity, pre-coordination level, extent of hierarchical and other interrelationships of the terms and the use of auxiliary precision decisions. A final layout of the thesaurus must be selected, whether purely alphabetical or includes systematic/classificatory sections.

4) Collect the Terms from Subject Literature, Users and Subject Specialists: Once the design and layout and the main subject groups are decided, the collection of terms for these subject areas may be taken up.

Background sources like descriptor lists or thesauri, subject headings lists, classification schemes, nomenclatures of single disciplines, treatises on the terminology of the concerned subject field, textbooks, subject dictionaries and encyclopedias, indexes of journals and abstracting periodicals, etc should be scanned for an initial list of the terms. It should be followed by scanning of subject literature. The abstracts and summary/ conclusions of the articles provide us a further list of terms. The users of the system and subject experts may be asked to list terms of importance in their subject fields. Their assistance and guidance would be more helpful.

5) Screening and Editing the Terms as per the Rules of Thesaurus Construction: Soergel suggests that the indexing would be most effective if done by a number of different subject experts in the same filed using terms of their choice. Though the compilers knowledge and familiarity with the subject field is an asset, it would be wise to obtain cooperation of user groups. The subject experts may be shown the lists of terms or draft classification schemes, in their own subject fields, and asked to comment, making amendments and adding terms. This will help to screen and edit the terminology. The Rules of Thesaurus Construction (For example, UNISIST Guidelines for the Establishment and Development of Monolingual Thesauri) as decided in the beginning should be followed consistently.

6). Recording the Terms in the Terms Cards / Thesaurus Form: Record the terms thus selected, screened and edited in the Thesaurus Cards/Forms. The cards designed for recording the terms are of 5"x 8" size. The information required for each term should include:

a) Index Term

b) Synonyms, near-synonyms and alternative word forms (UF)

c) Broader Term (BT)

d) Narrower Term (NT)

e) Related Terms (RTs), ie Non-hierarchically related terms

f) Source (if taken from dictionary, thesaurus, index, etc)

g) Scope notes and definitions (if necessary)

h) Broad Subject Groups/ Class Number

The details of the selected term and its relationships may be added gradually to the Thesaurus Form/Card during the process of compilation. A specimen of the Thesaurus Form/ Card is shown below:

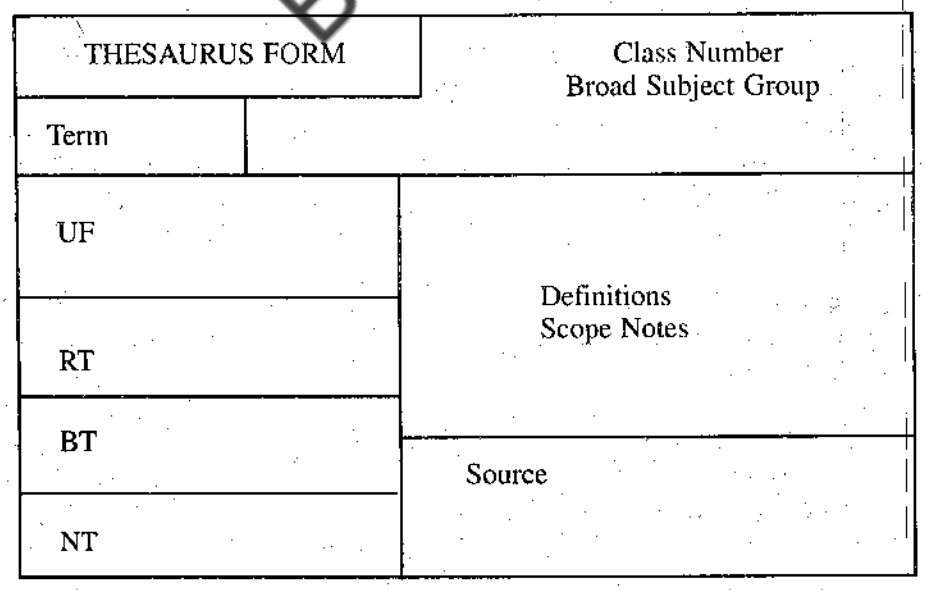


Figure-1: Thesaurus Form/Card
Source: Aitchison and Gilchrist (1972): *Thesaurus Construction*

7) Sorting and Grouping of Thesaurus Cards: All the thesaurus cards are to be sorted out and grouped according to their subject groups and sub-groups. Duplicate entries are to be eliminated in the preliminary scanning. If you have decided to have two parts (Alphabetical and Classificatory Parts), it is better to prepare two sets of cards as one set will be used for main part (Alphabetical Sequence) and the other one will be for the Systematic / Classificatory sequence). In fact, classificatory approach will speed up the preparation of hierarchical structures.

8) Prepare the Hierarchical Structure and other Associated Parts: From the groups of terms, interterm relationships are to be identified and hierarchical structures are to be developed among the descriptors. At this stage review of current literature on the subject and further consultations with subject experts help the compilers make them more clear about the synonymous terms and development of hierarchies. Facet analysis is another method which helps to identify and display of underlying structures.

9) Test the Thesaurus against a Selected Collection of Documents: Any thesaurus, thus compiled should be first tested against a selected collection of documents of the concerned subject field to examine its efficiency and use in information retrieval. It helps the compilers to evaluate their effort.

10) Get the Thesaurus is Evaluated by Subject Specialists and Users: The thesaurus is to be evaluated by subject specialists and a section of the users of the library and information system so as to get feedback on the working of the thesaurus. Based on the feedback the thesaurus could be refined.

8.5.2 Salton's Five Principles of Thesaurus Construction

Salton (as quoted by Townley and Gee, p.21) listed five principles of thesaurus construction, which should be engraved on the heart of any thesaurus compiler.

1) No very rare concepts should be included in the thesaurus since they could be expected to produce many matches between documents and search requests.

2) Very common high-frequency terms should also be excluded from the thesaurus since they produce too many matches for effective retrieval. (Individual high-frequency terms may be replaced by much more specific compound or hyphenated terms; for example, terms such as 'computer' and 'control' might well be eliminated in favour of a term such as computer control', since the former are clearly ambiguous in many contexts whereas the latter is much more specific).

3) Non-significant words should be studied carefully before they are included in the list of words to be eliminated.

4) Ambiguous terms should be included only for the senses that are likely to be present in the document collections to be treated. For example, at least two category numbers must be shown for the word 'field' corresponding on the one hand to the notion of subject area and on the other to its technical sense in algebra; however, no category need be shown to cover the notion of a patch of land if the dictionary deals with the mathematical sciences or related subjects.

(b) *Prepositional Phrases:* Ex: SCHOOLS FOR HANDICAPPED CHILDREN

The parts of a compound term which function as differences should be considered as potential sources of extra terms in a thesaurus. When a difference consists of an adjective, the noun from which the adjective was derived should be preferred as the extra candidate term. If these terms are accepted, the thesaurus should display reciprocal relationships between the extra term and the compound term as a whole. Example,

i) MARINE BIRDS
RT SEAS

SEAS
RT MARINE BIRDS

ii) SCHOOLS FOR HANDICAPPED CHILDREN
RT HANDICAPPED CHILDREN

HANDICAPPED CHILDREN
RT SCHOOLS FOR HANDICAPPED CHILDREN

2) Adjectives

Use of adjectives should be avoided as far as possible. However, they may be accepted as single words in situations where adjectives generally precede the nouns they modify and the user may be directed, on economic grounds, from a noun to an adjective. Example: France/ French, Sea/Marine, etc.

3) Adverbs

Adverbs such as 'Very' or 'Highly', should never be used alone as indexing terms. A phrase beginning with an adverb should not be accepted as an indexing term except when it has acquired a special meaning within a jargon. Ex: VERY HIGH FREQUENCY (VHF)

4) Verbs

Verbs expressed as infinitive or participles should not be used alone as indexing terms. Activities should be represented by nouns or verbal nouns. Ex: COOKERY (not 'Cook')

If the use of a noun to indicate an activity introduces a latent ambiguity, a qualifying phrase should be appended to each term.

Example: GOVERNMENT (activity)
GOVERNMENT (persons)

4) Abbreviations and Acronyms

Abbreviations and acronyms should not be used as preferred terms except when they are widely used and readily understood within the field covered by the thesaurus. Many acronyms and abbreviations are homographs, and the full form of the name should therefore function as the preferred term, with a reciprocal reference from the abbreviated form.

Example: WORLD HEALTH ORGANISATION
UF WHO
WHO *USE* WORLD HEALTH ORGANISATION

Abbreviations and acronyms may function as preferred terms if they have become so well established that the full form of the name is rarely used or is generally ignored. Reciprocal references should still be made between the full term and its abbreviation.

Example: UNESCO
UF United Nations Educational, Scientific and Cultural Organisation

United Nations Educational, Scientific and Cultural Organisation
USE UNESCO

5) *Choice of Singular or Plural Forms*

The decision to adopt singular or plural forms as indexing terms is likely to be affected by factors such as pre-coordinate/ post-coordinate indexing system used by the organisation and cultural factors. In order to avoid ambiguity where the singular form can refer to different concepts, one of which could be distinguished by expressing it in the plural. The terms can be divided into those that represent concrete entities, and those that refer to abstract concepts.

Nouns that represent concrete entities can be divided into two further categories:

(a) *Count Nouns*, ie names of countable objects that are subject to the question 'how many?' but not 'how much?'. These should be expressed as plural. Example: DOCUMENTS, POLITICAL PARTIES

(b) *Non-Count Nouns*, e.g., names of materials or substances which are subject to the question of 'How much ?', but not 'How many?'. These should be expressed as singulars. Eg: PAINT, QUARTZ, STEAM

The names of abstract concepts, e.g., abstract entities and phenomena, properties, systems of belief, activities and disciplines, should be expressed in their singular forms.

Examples:

Abstract entities and phenomena: PERSONALITY, WINTER
Properties: BRITTLENESS, OPACITY, SOLUBILITY
Systems of belief: MARXISM, SHINTOISM
Activities: IMMIGRATION, RESPIRATION
Disciplines: SOCIOLOGY, LIBRARY SCIENCE

7) *Homographs*

Homographs or polysemes (sometimes referred to by the broader term 'homonyms') are words with the same spelling but different meanings. When homographs are encountered, each term should be supplemented by a qualifying word or phrase.

Example: CRANES (birds)
CRANES (lifting equipment)

8) *Choice of Terms*

a) Spelling: The most widely accepted spelling of words should be adopted. A reference should be made from the variant spellings to the preferred form.

Example: Romania *USE* ROMANIA

b) Loan words and translations of loan words : If the loan word is more widely accepted, they should be incorporated If the translation becomes well-established, this should be preferred. Reciprocal references should be made between the preferred and non-preferred terms.

Example: X-RAYS *UF* Roentgen rays

c) Slang terms and Jargon : When a slang or jargon term emerges as an alternative to an existing and well established term, the established term should be chosen as the preferred term and the slang term should be admitted as a non-preferred term.

Example: ASSOCIATION FOOTBALL *UF* Soccer
Soccer *USE* ASSOCIATION FOOTBALL

d) Common names and trade names : A product is frequentlyknown by a widely recognised trade name. Where a suitable common name also exists, this should be adopted as the preferred term.

Example: POLYETHYLENE *UF* Polythene
Polythene *USE* POLYETHYLENE

e) Popular names and scientific names: The most likely to be sought by the users of the index should be preferred over the other form. 'Penguins' might be chosen as the preferred term in a general index and its scientific equivalent, 'Sphenisciformes' may be preferred in a zoological index.

f) Place names: The name preferred should be most familiar to the users of the thesaurus. Preference should be given to the official rather than the popular name.

Example: NETHERLANDS *UF* Holland
Holland *USE* NETHERLANDS

g) Proper names of institutions and persons: These terms are frequently excluded from a thesaurus, when included they should be recorded in their untranslated form. The names of international organisations should be expressed in their best-known forms.

9) *Compound Terms*

The establishment of procedures for dealing consistently with compound terms introduces one of the most difficult areas in the field of subject indexing. A general guideline of factoring

a compound term into separate components may not hold good in all circumstances. Therefore. certain terms have to be retained in its pre-coordinated form as per the context.

a) Semantic and Syntactical Factoring: There are two techniques for factoring (i.e., analysing a term into separate meaning elements) are recognised in indexing.

Semantic factoring: A term which represents a complex notion is re-expressed in the form of simpler or definitional elements, each of which can also be occur in other combinations to represent a range of different concepts.

Example: 'Thermometers' could be expressed by a combination of three terms:
TEMPERATURE & MEASUREMENT & INSTRUMENTS

This technique is not recommended as it leads to a loss of precision in retrieval. When a compound term has become so familiar in common use (eg: Data Processing), the proper names (eg: United Nations Organisation), and the terms in which the difference has lost its original meaning (eg: Lawn Tennis) etc. should be retained as compound terms.

Syntactical factoring : This technique is applied to compound terms, i.e., terms which are amenable to morphological analysis into separate components, each of which can be accepted as an indexing term in its own right. For example, 'Building Construction' can be expressed as 'Buildings' and 'Construction'.

Building Construction
USE BUILDINGS & CONSTRUCTION

8.6.2 Deciding on Interrelationships Among Terms

As you are learnt that a thesaurus is a controlled list terminology for a given subject area and shows relationships among the terms. It displays two types of relationships and they are - hierarchical and non-hierarchical. In hierarchical relationship, there are two types, namely, genus-species and part-whole. The non-hierarchical relationship can be further divided into equivalence and associative relationships.

These relationships can be shown through a chart as given below:

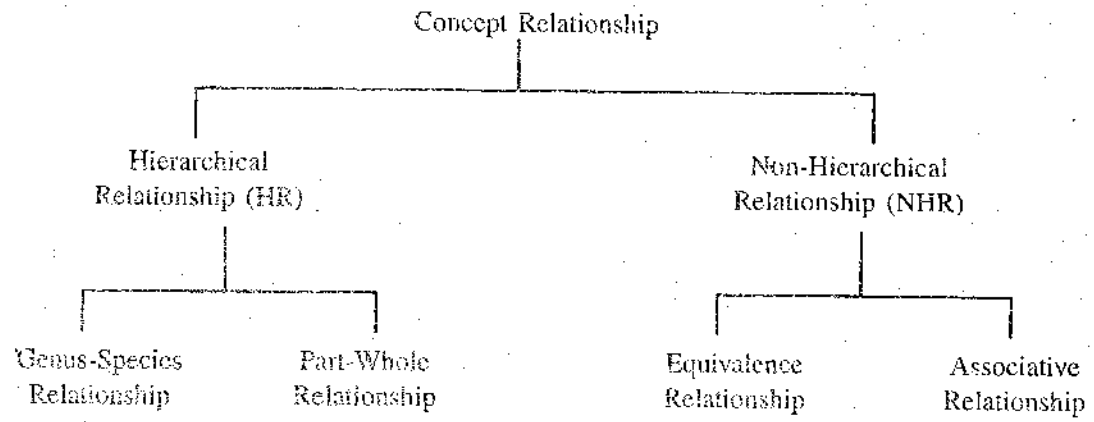


Figure-1: Concept Relationship in a Thesaurus

In constructing a thesaurus, it is necessary to understand various problems arising out of word forms in order to decide the above relationships. The compiler of a thesaurus needs to be aware of these problems, since they are likely to affect some of the decisions regarding the entries at a later stage.

I Equivalence Relationship

The equivalence relationship implies the control of the synonymy through preferred and non-preferred terms. The preferred terms are the terms used consistently to represent the concepts when indexing. Sometimes, these terms are also known as 'Descriptors', 'Keywords', 'Main terms', etc. The non-preferred terms are the synonyms or quasi-synonyms of terms, which are not assigned to documents, but which are provided as entry points in a thesaurus or alphabetical index. The user being directed by an instruction (i.e., USE or SEE) to the appropriate preferred term. Sometimes, these terms are known as 'Non-Descriptors'. These non-descriptors are also referred from the preferred terms by the instruction, i.e., UF (Used For). Let us examine some of these relationships in detail.

1) ***Synonyms:*** Synonyms are the terms whose meanings can be regarded as the same in a wide range of contexts, so that they are virtually interchangeable. Some of the more common classes of synonyms are given below:

(a) Terms of different linguistic origin
Examples: POLYGLOT; MULTILINGUAL

(b) Popular name and scientific name
Examples: ASPIRIN; ACETYLSALICYCLIC ACID

(c) Common nouns and trade names
Examples: VACUUM FLASKS; THERMOS FLASKS

(d) Variant name for emergent concepts
Examples: HOVERCRAFT; AIR CUSHION VEHICLES

(e) Current or favoured terms vs outdated or deprecated terms
Examples: DEVELOPING COUNTRIES; UNDERDEVELOPED COUNTRIES

(f) Variant spellings, including stem variants and irregular plurals
Examples: ROMANIA; RUMANIA; ROUMANIA

(g) Terms originating from different cultures sharing a common language
Examples: FLATS; APARTMENTS

(h) Abbreviations and full names
Examples: PVC; POLYVINYL CHLORIDE

(i) The factored and unfactored form of a compound term
Examples: COAL & MINING; COAL MINING

2) ***Quasi-Synonyms:*** These are the terms whose meanings are generally regarded as different in ordinary usage, but they are treated as though they are synonyms for indexing purposes. One user looking documents on one concept may also retrieve all documents on the other concept. Example: WETNESS; DRYNESS

The equivalence relationships are recorded in the thesaurus entries by choosing a term from the synonymous terms as a preferred term / descriptor and the others as non-preferred terms.

Example: DEVELOPING COUNTRIES
UF Underdeveloped Countries

Underdeveloped Countries
USE DEVELOPING COUNTRIES

II Hierarchical Relationships

The a priori relationships between concepts (e.g., as 'broader' and 'narrower') are of two kinds, namely, Genus-Species and Whole-Part Relationships. The earlier is universal in every information retrieval thesauri, while the later is, mostly, adopted for the thesauri of biological and medical sciences. The advantage of hierarchical order in thesauri over classification schemes lies in accommodating the polyhierarchy, when a descriptor has two or more broader terms.

Example: AIRCRAFT

By payload
FREIGHT AIRCRAFT
PASSENGER AIRCRAFT

By user
CIVIL AIRCRAFT
MILITARY AIRCRAFT

1) .The Generic Relationship: This relationship identifies the link between a class or category and its members or species.

If it is considered necessary, the generic relationship can be identified by the abbreviations BTG (Broader Term Generic) and NTG (Narrower Term Generic) or their equivalents in other languages.

Example: RATS
BTG RODENTS

RODENTS
NTG RATS

When a term belongs to a long and complex hierarchy it may be useful to indicate more than one level of subordination and/or superordination. In some thesauri the top term (TT) in

the hierarchy as well as the immediate broader term (BT) are also shown. Likewise, the broader term on first level of superordination, ie., next level (BT1); the broader term on second level of superordination, ie., higher level (BT2); the narrower term on first level of subordination, ie., next level (NT1); and the narrower term on second level of subordination, ie., lower level (NT2); are also shown in the thesaurus entries.

2) The Part-Whole Relationships: This relationship covers a limited range of situations where the name of a part implies the name of its possessing whole in any context. The terms can then be organised as a hierarchy, the name of the whole serving as the superordinate term, and the name of the part as the subordinate term. This applies to four main classes of terms.

(a) Systems and organs of the body:

Example:
CIRCULATORY SYSTEM
CARDIO-VASCULAR SYSTEM
VASCULAR SYSTEM
ARTERIES
VEINS

If considered necessary, the partitive relationship can be shown with the abbreviations BTP (Broader Term Partitive) and NTP (Narrower Term Partitive) as shown below:

CARDIO-VASCULAR SYSTEM

BTP CIRCULATORY SYSTEM
CIRCULATORY SYSTEM

NTP CARDIO-VASCULAR SYSTEM

(b) Geographical locations:

Example:
INDIA
ANDHRA PRADESH
HYDERABAD
VISAKHAPATNAM

c) Disciplines or fields of discourse:

Example:
SCIENCE
BIOLOGY
BOTANY
ZOOLOGY

(d) Hierarchical social structures:

Example:
ARMIES
CORPS
DIVISIONS
BATTALIONS
REGIMENTS

III Associative Relationships

The association relationships are the non-hierarchical relationships among the preferred terms, which are of diversified nature in various thesauri. The preferred terms are related to each of the associated terms (RTs ie., Related Terms) in an unspecified manner, but shown primarily to suggest other approaches that might be taken in conducting a search. A. Neelameghan (1975) has identified and enumerated twenty-nine different types of non-hierarchical relationships, while T.N.Rajan's study on Thesaurofacet has categorised the non-hierarchical relationships into five types. However, the RTs function in the same way as alphabetical indexes to classification schemes in displaying the 'Distributed Relatives'.

Two kinds of terms can be linked by the associative relationship: (a) those that belong to the same category, and (b) those belonging to different categories.

(a) Terms belonging to the same category: A user seeking documents on one of the terms should be reminded of the other terms. These links are shown only in the alphabetical parts and are not indicated in the hierarchies.

Example:	SHIPS	BOATS
	BT VEHICLES	*BT* VEHICLES
	RT BOATS	*RT* SHIPS

(b) Terms belonging to different categories: Terms belonging to different categories can be associated with regard to situation, operation or action to which both of them belong.

i) A discipline or field of study and the objects or phenomena studied:

AESTHETICS	BEAUTY
RT BEAUTY	*RT* AESTHETICS

ii) An operation or process and its agent or instrument:

DATA PROCESSING	COMPUTER SYSTEMS
RT COMPUTER SYSTEMS	*RT* DATA PROCESSING

iii) An action and the product of the action:

WEAVING	CLOTH
RT CLOTH	*RT* WEAVING

iv) An action and its patient:

IMPRISONMENT	PRISONERS
RT PRISONERS	*RT* IMPRISONMENT

v) *Concepts related to their properties:*

POISONS	TOXICITY
RT TOXICITY	*RT* POISONS

vi) *Concepts related to their origins:*

INDIA	INDIANS
RT INDIANS	*RT* INDIA

vii) *Concepts linked by causal dependence:*

DISEASES	PATHOGENS
RT PATHOGENS	*RT* DISEASES

viii) *A thing and its counter agent:*

PESTS	PESTICIDES
RT PESTICIDES	*RT* PESTS

viii) *Syncategorematic phrases and their embedded nouns:*

MODEL SHIPS	SHIPS
RT SHIPS	*RT* MODEL SHIPS

8.7 EXAMPLES OF THESAURI

A number of libraries and information centres focussed their attention in designing information retrieval thesauri in the 1960s and 70s. While some LICs developed their own thesauri for their limited use, some organisations developed thesauri and made them available in published form in varying levels of subject specialisation. The pattern of structure and format differ from thesaurus to thesaurus. Let us examine the salient features of major information retrieval thesauri.

1) Thesaurus of Engineering and Scientific Terms (TEST)

The TEST is known to be the largest thesaurus, both in coverage and use, so far developed. It was compiled with the continuous efforts of a team of over 300 scientists and engineers and published in December 1967. It is originally based on the *Thesaurus of Engineering Terms,* published in 1964 by the Engineers Joint Council (USA).

The TEST includes 17,180 descriptors or index terms with 5,554 lead-in terms or references. It shows inter-term relationships, such broad terms, narrow terms and related terms. In addition to Alphabetical Part (main part), it has one Hierarchical Index, a Permuted Index and a Subject Category Index. It also displays the numbers against each descriptor, indicating the COSATI (Committee on Scientific and Technical Information) Subject Category fields and groups.

The Hierarchical Index displays two or more levels of narrower terms (lower links). The Permuted Index provides additional approach points (lead-ins) to the individual terms of the compound terms. The Subject Category Index helps as classificatory aid and it is based on COSATI Subject Category List.

2) *Thesaurofacet*

The *Thesaurofacet, a thesaurus and faceted Classification for Engineering and related Subjects* was published by the English Electric Company (EEC) in 1969. It contains 16000 index terms with 7000 additional lead-in terms or references. As the subtitle indicates, the scope of the thesaurus is mainly the engineering and related subjects of interest to engineers.

As discussed in Section 8.4.1 Types of Thesauri, a thesaurofacet is a thesaurus integrated with a classification scheme. The *Thesaurofacet* used the fourth edition of English Electric's *Classification for Engineering*. The classification part of the *Thesaurofacet* can be used independently as a library classification for shelf arrangement and for other purposes. The Thesaurus part shows the relationships of concepts through the BT-NT-RT structure. Class numbers are indicated against the descriptors in the thesaurus part. Some of the RTs or NTs are shown with an 'A' in brackets indicating the position in the additional hierarchy. The class numbers derived by synthesis are shown by the use of *synth*, and an *S* preceding the constituent terms.

3) *Information Retrieval Thesaurus of Education Terms*

Information Retrieval Thesaurus of Education Terms of the Case Western Reserve University was published in 1968. The thesaurus is organised into three inter-related parts: 1) Alphabetical Array, 2) Faceted Array, and 3) Permuted List of Descriptors. In the Thesaurus Part, the display of terms include Used For terms, Narrow Terms and Related Terms. An elaborate Scope Note (SN) is also included under each descriptor. Numerical RTs are the special feature of the Thesaurus, which direct the users from the descriptor to the relevant part of the Faceted Array (a classified display). The Thesaurus later became the base for the famous *Thesaurus of ERIC Descriptors*.

4) *Roots Thesaurus*

Roots Thesaurus was published by the British Standards Institution. It is based on Roget's original principle of systematic list, accompanied by an alphabetical display. It is very difficult to see any fundamental difference between the *Roots Thesaurus* and *Thesaurofacet*, as far as the structure is concerned. The main sequence is systematically arranged using an upper case letter notation. Notation is semi-hierarchical. It uses mathematical symbols: < (Broad Term), > (Narrow Term), — (Related Term), *< (BT in alternative hierarchy), = (Non-preferred Term/ Synonym), —> (USE), + (Term used to synthesize a given concept), ** (Synthetised Term), [...] Scope Note, (By...) Facet indicating the structure of the main sequence not used for indexing purposes, etc. The Roots Thesaurus is intended to be a computer-based multilingual system. French translation is available.

5) *OECD Macrothesaurus*

OECD Macrothesaurus: a basic list of economic and social development terms. (English edition) was published in 1972. It is the revised edition of the *Aligned List of Descriptors*, a multilingual compilation, published in English, French, German and Spanish. It was compiled by various important international organisations and the updation is done by UN Family. The headquarters of the Macrothesaurus is based at the International Labour Organisation (ILO).

The *Macrothesaurus* is a bilingual compilation covering the subject fields of economics and social development. It is alphabetically arranged, taking English terms into consideration. There are two parts: 1) Alphabetical Thesaurus, and 2) Thesaurus by subject fields. Both the parts provide BT-NT-RT structure. The second part provides a hierarchical display of terms.

6) *Unesco Thesaurus*

Unesco Thesaurus: A Structured List of Descriptors for Indexing and Retrieving Literature in the fields of Education, Sciences, Social Sciences, Culture and Communication was compiled by Jean Aitchison and published by Unesco in 1977. It has been designed as a working tool of the Computerised Documentation System of Unesco. It is used in the formation of CDS database and for retrospective searching and SDI services.

The pattern of structure of the *Unesco Thesaurus* to some extent follows Thesaurofacet. The compilation is in two volumes: 1) Introduction, the Classification, Thesaurus, the Permuted Index and Hierarchical Display of Terms; and 2) Alphabetical Part. The Thesaurus contains 8,500 descriptors.

7) *INIS Thesaurus*

The INIS Thesaurus was published by International Atomic Energy Agency, Vienna, 1973. It is based on the earlier thesaurus, namely, the *Euratom Thesaurus*. It was compiled specifically for the decentralised preparation of input to the INIS.

The unique feature of the *INIS Thesaurus* is the use of 'terminological charts'. The purpose of the terminological charts is to display the descriptors in the context of their hierarchical and other semantic relationships. The terms are grouped into clusters. Under the broadest term of each cluster, the other terms are arranged in smaller boxes. The clusters re connected by lines of various thickness, showing the "see also" and "related term" cross references. The Thesaurus follows the usual BT-NT-RT pattern in its structure, however the indication of the levels of BTs and NTs by indention (Eg: BT1, BT2 and BT3; NT1, NT2 and NT3) in the alphabetical display of terms helps in identifying the importance of the descriptor in the subject.

8) *Medical Subject Headings (MeSH)*

Though the title is fashioned as a subject headings list, *MeSH* is a thesaurus. It is published as Part-2 of *Index Medicus* in January each year. It is used in the computerised services of *MEDLARS*, where it is possible to prepare specific search formulation by co-ordinating a number of terms and other tags. An annotated version of *MeSH* is also published for use in *MEDLINE*.

The display pattern of the inter-term relationships in the *MeSH* differs from the usual BT-NT-RT nomenclature. The main device used for showing relationships is a full hierarchical classification scheme. Each descriptor in the Alphabetical Part carries a notational tag and a category number. The category numbers are given in parenthesis after the terms. These numbers direct the user to the Category List. Category Lists are hierarchical displays or 'Tree Structures'. In some cases a term may occur under more than one category and hence, two or more numbers are found in the parenthesis. XU and XR indicate 'see under' and 'see also related' (headings) respectively. 'x' mark indicates that this is a minor descriptor, it is used as an indexing term in *MEDLARS*, but not in *Index Medicus*.

The annotated version of *MeSH* contains the headings with fuller scope notes, historical notes and search notes. The 'new headings' replacing the 'previously indexed under headings' are given in every publication.

9) *Other Information Retrieval Thesauri*

There are a number of thesauri published either by the organisations or individuals. Some of them are listed below:

Spines Thesaurus (published by UNESCO under UNISIST project)
INSPEC Thesaurus (Institution of Electrical Engineers, London)
EUDISED Multilingual Thesaurus
CIS Thesaurus of International Occupational Safety & Health (ILO, Geneva)

8.8 LET US SUM UP

Let us recapitulate what has been discussed in this unit.

* Ther term 'Thesaurus' was introduced in 1852 by Peter Mark Roget through his dictionary entitled, Thesaurus of English Words and Phrases.

* Thesaurus entered LIS field in the early 1950s through Helen Brownson paper at the Dorking Conference and H.P. Luhn's 'Word Association Matrix'.

* Thesaurus is a controlled and dynamic vocabulary of semantically and generically related terms which covers a specific domain of knowledge.

* There are three kinds of relationships are shown in a thesaurus. They are hierarchical, equivalence and associative relationships.

* Thesaurus is used as an effective indexing technique to improve the precision and recall in information retrieval systems.

* UNISIST Guidelines for Establishment and Development of Monolingual Thesauri recommends the methods of treatment of word forms and construction of thesauri.

8.9 REFERENCES AND FURTHER READING

AITCHISON, J. and A. Gilchrist. *Thesaurus Construction: a practical manual.* London: Aslib, 1972.

Encyclopedia of Library and Information Science/ edited by Allen Kent, et al. New York: Marcel Dekker, 1980. Vol.30; p.416-463.

FOSKET, A.C. The Subject Approach to Information, 4th ed. London: Clive Bingley, 1982.

FOSKETT, Douglas J. "Thesaurus". IN *Encyclopedia of library and information science/*edited by Allen Kent, Harold Lancour and Jay E.Daily, New York: Marcel Dekker, 1980. Vol.30; p.416-463.

GILCHRIST, Alan. *The Thesaurus in Retrieval.* London: Aslib, 1971.

GUHA, B. *Documentation and information: Services, techniques and systems,* 2nd rev ed. Calcutta: The World Press, 1983.

GUIDELINES for the establishment and development of monolingual thesauri / PGI and UNISIST. 2nd rev. ed. Paris: Unesco, 1981.

HUTCHINS, W.J. *Languages of indexing and classification: a linguistic study of structure and functions.* Herts: Peter Peregrinus, 1975.

INDIRA Gandhi National Open University. *Information Processing and Retrieval* (MLIS-3; Block-1: Unit-4: Thesaurus). New Delhi: IGNOU, 1995.

NEELAMEGHAN, A. and R.Maitra. "Non-hierarchical associative relationships among concepts: Identification and typology". *IN FID/CR Report* No.18; 1978.

RAIZADA, A.S. et al. "Principles of thesaurus construction and Indian Science Abstracts indexing system". IN *SEMINAR on Thesaurus in Information Retrieval* (Bangalore : 1975), jointly organised by DRTC and INSDOC. Bangalore: DRTC, 1975. (Paper AD)

RAVICHANDRARAO, I.K. "Semi-automatic construction of thesaurus". ibid! (Paper BB)

SEETHARAMA, S. "Semi-automatic construction of thesaurus". *IN DRTC Workshop on Information Retrieval.* Bangalore: DRTC, 1992.

SEVENDIUS, e. "Design of controlled vocabularies in the context of emerging technologies". Library Science 25(4); 1988

TONLEY, Helen M. and Ralph D..Gee. *Thesaurus making.* London: Andre Deutsch, 1980.

VICKERY, B.C. *Techniques of information retrieval.* London: Butterworths, 1970.

8.10 ASSIGNMENT

1) Select a small subject area in which your organisation is interested and compile a microthesaurus. Test it on a selected collection of documents relating to that subject field.

2) Compare any two thesauri available in your organisation or accessible to you.

8.11 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) What is a thesaurus ? Describe the structure of a thesaurus entry with an example.

2) Discuss the methods of treating the terms in a thesaurus as recommended by UNISIST Guidelines for establishing and development of monolingual thesauri.

3) Explain in detail the method of constructing a thesaurus.

II SHORT NOTES

a) Syntactic and Semantic factoring

b) Salton's Principles

c) Thesaurus Form

BLOCK - III : INFORMATION STORAGE AND RETRIEVAL (ISAR) SYSTEMS

The Information Storage and Retrieval (ISAR) Systems are concerned with two important components - Storage and Retrieval. These two activities are interdependent as retrieval is dependent on how data/information is stored. The ISAR is a process concerned with the selection, representation, storage, organisation and accessing of information items to meet the specific requirements of a user community. The basic components that are integrated into an ISAR system are information content, utility of resources, users, documentary resources, performance resources and economies. The designers of the ISAR systems need to consider three basic activities, i.e., information resource building, database creation and maintenance, and information retrieval and dissemination.

File organisation is one of the important aspects of an ISAR system. In library and information centres, the data stored about the documents is descriptive and exhibit specific characteristics of repeatable fields and many fields are with varying length. These diffeerences in file and record organisation play a crucial role in maintenance and retrieval process. Some database management system (DBMS) applications for integrated library automation are designed and developed using variable length record descriptions.

Evaluation of ISAR systems is carried out in terms of Systems effectiveness evaluation, Cost-effectiveness evaluation and Cost-benefit evaluation. The methods of Systems Effectiveness evaluation concentrate on the user requirements, Coverage, Recall, Precision, Response time, User effort, Form of output, etc. are the basic elements in understanding the system effectiveness. However, a great deal of effort has been made to develop other means of evaluation, i.e., fallout, selectivity, specificity, novelty, noise, etc.

There have been several experiments and case studies undertaken towards evaluation of ISAR systems. The first experiment, namely, CRANEFIELD I, compared the efficiency of four indexing systems and associated file organisations in terms of Recall and Precision ratios. In CRANEFIELD II, the components of indexing language and the effects of various components on overall system effectiveness were studied. Later various case studies have been undertaken for the evaluation of operational systems (FAIRS and MEDLARS), evaluation of experimental system (SMART) and evaluation of expert system based user interface (MEDLINE-CANSEARCH) and evaluation of the effectiveness of user interface, which have been discussed in the last part of this block.

In the present block, there are four units, viz.,

Unit- 9	:	Information Storage and Retrieval (ISAR) Systems
Unit-10	:	File Organisation in ISAR Systems
Unit-11	:	Evaluation of ISAR Systems - Methodology
Unit-12	:	Evaluation of ISAR Systems - Experiments and Case Studies

UNIT-9 : INFORMATION STORAGE AND RETRIEVAL *(ISAR)* SYSTEM - AN OVERVIEW

Structure

9.0 AIMS AND OBJECTIVES

The unit aims to provide an overview of an Information Storage and Retrieval (ISAR) System, including its objectives, types and various components..

After studying the unit you could be able to

- list out various types of information systems
- define information storage and retrieval systems
- explain objectives of an ISAR system
- describe various components of an ISAR system.

9.1 INTRODUCTION

Traditionally, librarians are concerned with document-based storage and retrieval systems, which use the basic tools such as classification and cataloguing. These tools have been refined over the years, but only to handle the macrodocuments.

As you have studied in the previous Block of this Course, the indexing systems were developed and used as retrieval systems even before the 1940s. They were purely manual. They are pre-coordinated and linearly organised systems and therefore, non-manipulative. They provided very little flexibility in searching operations. In the 1940s and after period, there were some major developments in indexing systems, namely, post-coordinated systems. These systems are flexible and manipulative. Eg: optical coincidence and uniterm systems.

The explosion of literature in the form of microdocuments on the one hand and the growing number of users demanding more specialised literature/information have led to information scientists to depend on advances in technology and the techniques associated with it. Computerised database creation and the sophistication in indexing techniques have eased the problems of storing and handling of large volume of data/ information and exploded the opportunity for the realisation of retrieval systems with greatly enhanced capabilities. Therefore, the focus of the information scientists for the recent past few decades is on the design and development of more powerful Information Storage and Retrieval (ISAR) systems.

Computer-based information systems came into being in the mid-1960s. These were off-line and used batch process systems and basically tape oriented. The era of on-line interactive information systems started in the late 1970s.

9.2 INFORMATION SYSTEMS

Computer-based information systems can be categorized into:

a) Information Storage and Retrieval (ISAR) Systems,

b) Database Management Systems (DBMS),

c) Management Information Systems (MIS),

d) Decision Support Systems (DSS), and

e) Question-Answering Systems (QAS).

All these systems exhibit similarities to some extent in the area of information processing but differ in their functionalities.

9.2.1 Information Storage and Retrieval Systems

Many of the items found in an Information Storage and Retrieval (ISAR) system are characterized by an emphasis on narrative information i.e. information being processed consists of documents. In this context, information retrieval deals with the representation, storage, organization and access to documents or representatives of documents (documents surrogates e.g. bibliographic references). The input information is likely to include the natural language text of the documents or of documents excerpts or abstracts. A query put to the system usually yields a set of references which are intended to provide the user with information about the items of potential interest. The usefulness of an information storage and retrieval system depends crucially on currency and completeness. New items are constantly added to the collection to maintain currency, and the collection contains a large proportion of the items of potential interest for completeness. This implies that the data stored in an information storage and retrieval system is basically static i.e. not supposed to change and hardly any item is removed from the system. For example, in a system that captures bibliographic references, an entry is made of a document and time is not going to change the values stored i.e. the author, title, publisher and other details will not change. In information storage and retrieval systems, fast retrieval of information predominates data redundancy control or disc space conservation. This objective is achieved by keeping the information about an item included in the system at one place i.e. in a single record and providing multiple access points to this item through indexes.

9.2.2 Database Management Systems

Database Management Systems (DBMS) are concerned with the storage, maintenance, and retrieval of data facts available in the system in the explicit form. That is the information does not appear as natural language text but is available in the form of specific data elements stored in tables. In a database environment, each item or record, is thus separated into several fields, and each field contains the value for a specific characteristic or attribute identifying the corresponding record. Specific values of these characteristics or attributes are used as identifiers for the individual records. For example in an information storage and retrieval system dealing with bibliographic descriptions, affiliation of an author may be included in a single field as a continuous text, but still is searchable through . specific indexing techniques adopted by these systems. But in a DBMS environment the same affiliation field is decomposed into multiple data elements such as Department name, Institute name, Street, City, country and PIN code etc. In contrast to information storage and retrieval systems, the data is dynamic i.e. supposed to change frequently (e.g. circulation system in a library where the status of a document changes frequently as 'in issue' or 'on stack').

The processes involved in a database management system include the storage and retrieval of data, the updating or deletion of data, the protection of data from unintentional or deliberate damage or misuse, and the transmission of data to remote users or other data management systems. The output of the system may consist individual records, portions of records, tables, or other arrangements of the data (not set of references). Each search request must state the specific values of identifiers for the records of interest and the retrieved information will include all records which match the stated search request exactly. For example, in banking operations a search for a specific consumer should exactly match with the database record before a credit or debit transaction takes place. As data privacy and precise record retrieval predominate the DBMS applications, the software provides several security mechanisms at database, record, field and process levels and several standardization principles and data integrity checks for data validation. The information contained in database management systems include a good deal of numerical data, and statistical and computational facilities are also provided to manipulate the numbers.

9.2.3 Management Information Systems

In an organization usually different databases are created and maintained to meet the specific data processing requirements of the departments i.e., individual data files were created and different DBMS are applied, often running on different computers, to meet the needs of various business activities such as accounting, personnel, marketing, and production. As these systems proliferated, there developed a growing realization on management's part that, if properly integrated, some of the data gathered to support functionally oriented clerical processing could be used to support management decision making. the notion of gathering and processing data from multiple sources to make them useful for decision making is fundamental to Management Information Systems (MIS) concept. Products of a management information system are usually prespecified outputs, either in the form of printed reports or video terminal presentations. In either case they will have been designed before they are actually needed. Activities of both the operating and middle management levels are likely to be supported by the management information systems. In this context a management information system is a DBMS tailored to the needs of managers. The functions performed by a manager in a given corporation depend on the availability of many kinds of data. Information leading to the choice of possible alternatives by the manager presented in terms of ranges of values of a particular attribute is of particular interest. In management information systems , the information is subjected to special processing not normally available in the database management systems.

E.g.: Inventory replenishment decisions and inventory policy decisions would be supported by reports describing stock levels and product demand: information developed from data drawn from order entry, material management, and marketing resources.

9.2.4 Decision Support Systems

The operating management activities supported by DBMS in a specific area and the pre-specified decision tasks supported by a management information system provide two examples of structured decision activity. This is the easiest problem to approach with a computer system because the outputs and inputs can be determined, designed and tested before the system is

implemented. Many of the decisions made at middle management level and most of the decisions made by the upper management are less structured.

Unstructured decisions tend not to be repetitive: decisions based on the same set of facts and analysis procedures are not likely to be made on periodic basis. Decisions to manufacture a new product and capital investment decisions are likely to be in this category. In fact, these high level decisions such as choosing where to locate a new plant may be made only once or a few times in the life time of an organization.

The input of management and staff knowledge, insight, intuition, and specialized expertise is necessary to make these major decisions that fall into the unstructured category. In these systems a single cooperating structure that includes information retrieval systems, database management systems, statistical analysis systems, computer graphics systems and other technical capabilities like modeling (simulation models, models of operations research), collectively provides a friendly interface in support of the decision making process. Decision support systems exist on a limited basis for narrow ranges of users employing databases in restricted subject areas.

9.2.5 Question-Answering Systems

Thse systems provide access to factual information in a natural language setting. The stored file often consists of large number of facts relating to special areas of interest, together with general world knowledge covering the context within which conversation between persons usually takes place. User questions may be received in natural language form, and the system responses may also be furnished in natural language formulation. The task of the question-answering system consists in analyzing the user query, comparing the analyzed query with stored knowledge, and assembling a suitable response from the apparently relevant facts.

9.3 INFORMATION STORAGE AND RETRIEVAL (ISAR) SYSTEM - OBJECTIVES AND FUNCTIONS

As the name indicates ISAR systems are concerned with two important aspects: storage and retrieval. These two activities are interdependent and inseparable as retrieval is dependent on how the data/information is stored. Information retrieval is an activity interposed between a potential user of information and the information collection itself. The information (retrieval) system captures the wanted items and filters out unwanted items.

9.3.1 Objectives and Functions of an ISAR System

The aim of an ISAR system is to provide right information at the right time in a right form as desired by the information seeker. This is not a simple task. The end-users are not aware of the problems and complexities of information access, storing/organising and retrieving it. The designers of information systems take considerable amount of difficulty in making the system readily accessible to the seekers of information. According to M.L. Pao (1989), for service oriented systems, user's input is an important consideration to be incorporated in setting

he overall objective for the system. There may be several classes of objectives. They should be anked according to priorities. The objectives are also associated with the functions of an information retrieval systems. Pao (1989) identified the following types of objectives to be considered in relation to information retrieval system:

* Information content of information resources collected
* Utility of information resources
* Users
* Documentary resources
* Performance resources
* Economies

The Information Content of an information retrieval system may be subject oriented, mission-oriented, multidisciplinary, or interdisciplinary. For example, the National Medical Library is concerned with all aspects of medical sciences. A database produced by a rural development institute concerns with several subject areas, such as economic, sociology, public administration, education, health, technology, etc. In fact, information retrieval systems exist to provide information resources to the users.

Utility of the information records collected is the primary concern of any information system. To promote on-going research in the subject area, the system has to acquire and provide potentially useful information to users on a continuous basis so that they are kept up-to-date in their subject fields. This serves current awareness function. The indexing and abstracting services have different kind of utility in locating the bibliographic references relevant to users' interest. Online information systems or CD-ROM databases serve in providing information instantly.

Identification of Users or user groups is the most obvious objective that needs to be defined. Often the intended user groups may vary with that of actual user groups. For example, the intended users of a medical information system are clinicians and biomedical researchers, but it is used by the librarians (as intermediary), biochemists and other health professionals. An IR system should have a specific user groups.

After identification of information resources, their utility and user groups, IR system needs to determine the *documentary resources*. It also includes the depth of subject coverage, type of documents, years of coverage, languages, etc. of the documents. Users' needs, storage limitations, financial resources, personnel requirements, etc. often impose certain restrictions on the decisions taken by the designers of IR systems.

Performance criteria is the key to design of an information retrieval system. Effectiveness of the system depends on completeness, subject coverage and relevance of the material included in the system. Most users do not know what to expect of an information retrieval system. There are no commonly agreed upon criteria, however, relevance of the information, speed of service, and unit cost associated with it are cited as performance criteria.

Economics is often taken as the determining factor in many IR system designs. However, it should not be the sole determinant factor. Most of the online information systems in the beginning of their establishment subsidised their services as users could not afford the costs associated with it. By the 1980s, these online systems have become commercially viable and profitable systems. Today, most systems are operated by profit-making organisations also reflect the increasing societal values placed on information.

9.3.2 Records in IR Systems

The records in a retrieval system can be of several kinds:

1) Quantitative or Qualitative data about variables of interest (e.g. statistical data on population, food production etc.)

2) Texts (including illustrations) on every kind of subject

3) Drawings, graphs, charts, maps and other kinds of graphic material (e.g. engineerin drawings, Geographic Information Systems (GIS)

4) Computer programs

5) Description of objects - for example of minerals, laboratory apparatus, industrial equipment etc. (e.g. brochures on different types of equipment)

6) Names and locations - of people, institutes, manufacturers etc. (e.g. Telephone directory, Directory of manufacturers etc.)

7) Bibliographic references - i.e. indicators of the identity and location of documents (e.g. Abstracting and indexing journals, OPAC)

These entities are represented in an information system with the application of a suitable software. The term 'entity' is widely used in information retrieval systems to mean anydistinguishable object or concept that is to be represented in the system. The storage, organization and access mechanisms also vary according to the type of entity that is represented in 'an information system.

9.3.3 Knowledge and Skills required for Designers of IR Systems

Information retrieval is a process concerned with the selection, representation, storage, organization, and accessing of information items to meet the specific requirements of a user community. Therefore, the producers of databases study what type of information is generated and how to collect and organise it to satisfy the users needs. The information scientists working with the database creation play a crucial role in designing the ISAR systems. They need to possess certain qualities and skills. The essential skills and knowledge required by an information scientist are - 1) Knowledge of the subject they are handling, 2) Current developments and trends in subject literature, 3) Knoledge of the vocabulary control systems, 4) Users' needs and their information seeking behaviour, 5) Methodology of database creation and representation of information in them, 6) Methods of evaluating the ISAR systems.

9.3. 4 Factors Influencing the Design of an ISAR System

Any information storage and retrieval system is designed and developed to meet the specific requirements of a user community. The factors that influence the system design are grouped into external and internal.

The external factors include:

* the requirements of the user community,
* the types of documents that can meet these requirements
* the sources of data - in-house generated, imported from external sources or both
* exchangeability - serves in-house users only
  - exchanges data with other organizations using same ISR system
  - exchanges data with other organizations using different ISR system

If the proposed system is expected to import and export data, conversion programs that can map data elements of external source into data elements of the internal source and vice versa are required.

The internal factors include:

* kinds of documents to be included (e.g. monographs, journal articles, theses, reports, patents etc.)
* levels of descriptions of these selected documents (e.g. monographic or analytical level)
* Data elements to be included for the description of the above types of documents at specific levels as per international standards
* skills available in the institute to design and develop the system
* users of the system and their skill levels i.e. meant for library staff or end users
* Searchability -Boolean, truncated, free text etc.

9.4 TYPES OF *ISAR* SYSTEMS

Information storage and retrieval systems may be grouped into three types based on the content of the system: a) Reference retrieval systems, b) Document retrieval systems, and c) Fact retrieval systems.

9.4.1 Reference Retrieval Systems

In a Reference Retrieval System, a record is made for each document of interest with brief descriptions of the document and location details. For example, an article published in a

journal is included with author, title; and abstract details along with the journal name, volume, issue, pages and year of publication that help to locate the article. A search of the system results in a set of references to the documents and not the document itself. Good examples are abstracting and indexing systems. These systems heavily depend on secondary indexes to provide multiple access points to the records in the system (e.g. to search the system by author, title, or subject terms).

9.4.2 Data Retrieval System

In a Document Retrieval System, the full text of the document is stored along with illustrations, tables, graphs etc. using specific software. A search on this system with the aid of indexes retrieves the total document usually in the form of images (e.g. journals published in electronic form, computer-output microforms).

9.4.3 Fact Retrieval System

The third kind of retrieval system is Fact Retrieval System, where the details of an item of interest are stored in a specified format. For example; the directory of publishers includes data about the name, place, address and other details of each publisher. Similarly, a file containing the properties of chemical elements may include information on atomic number, atomic weight and other properties of the element. A search on these system retrieve data or facts about the items included in the system. It is also possible to search the system on a combination of properties.

9.5 COMPONENTS OF AN *ISAR* SYSTEM

The three basic activities of an information storage and retrieval system are information resouce building, database creation and maintenance, and information retrieval and dissemination. To perform these activities in an efficient manner, the system should have an integrated set of components (modules): a document selection subsystem; a data input and validation subsystem; an indexing subsystem; a vocabulary control sub-system; a search sub-system for interactive and batch mode operations; a user interface that helps the novice user in searching; an output/report generation subsystem; and a system usage monitoring subsystem.

9.5.1 Document Selection Sub-System

Documents/records may be selected manually keeping in view the scope of the system. If the system is supposed to import data from external sources, a profile with necessary search strategies is created and executed at regular intervals to extract the relevant records. The data elements of records extracted from an external source are to be mapped into the data elements of the internal record structure and conversion programs are to be run before loading.

9.5.2 Data Input and Validation Sub-System

Data may be entered in to the system in interactive or batch modes. Interactive data entry is carried out with the help of data entry worksheets, while specific utilities are used

to import data in batch mode. A set of rules may be embedded in to data definition so that all the records are validated for the presence of mandatory fields, and optional fields based on the type of document and the level of description before being loaded into the system. For example, the validation procedures may check for the presence conference name, place and date if the document belongs to a conference; for the presence of Degree if the document belongs to a theses and so on. A set of authority files may also be used to validate the data entered.

9.5.3 Indexing Sub-System

The indexing sub-system provides one or more options to the designer such as: indexing entire fields, subfields, keys generated from multiple fields, significant words in a continuous text, indexing numerical values etc. Some software make provision either to select a new field for indexing or drop a field from the index with out disturbing the existing index. The system may also support soundex and plural control.The options used in indexing a field plays a major role while retrieval.

9.5.4 Vocabulary Control Sub-System

An on-line thesaurus is used as an aid during indexing and retrieval operations. The records having valid descriptors are accepted by the system and others are rejected. The software may also provide the facilities to index the records on preferred terms only, by switching automatically from user entered non-preferred terms. Similarly, the system may switch from non-preferred terms entered by the end user during a search process to preferred terms. The browse facility of the thesaurus may provide the user with a set of broad, narrower, synonymous, and related terms for a term that helps the user to form his search strategy in precise manner.

9.5.5 Search Sub-System

The search sub-system supports:

* multiple key searching with Boolean operators
* Free text searching i.e. specific word/ word combinations; proximity/adjacency or phrase like searching
* Wild card searching e.g. An*gonism
* Truncated searching - left/right truncations
* Hierarchical searching with the aid of thesaurus
* Numerical searching e.g. dates and numerical ranges
* browsing indexes and thesaurus
* search refinement by combining sets generated by any one of the above options

* selecting the most relevant records from a set
* sorting the records in a set on one or more keys
* profile based searching in batch mode

The systems that support interactive searching are categorized into: command-driven systems, menu-driven systems, GUI (Graphical User Interfaces) based systems.

Command-driven systems depend on the user's memory i.e. the user should know exactly the command to be used and the syntax of the command along with necessary parameters. Any slight variation may result in an error message. These systems provide the user with a prompt to which a response is made by a command line with necessary parameters or a file name with user defined commands.

In menu-driven systems, text menus with a list of options are provided and the user selects one of the options either by typing a number or character representing the option, or by simply pressing the Enter key when a particular option is highlighted. The selection of an option may lead to an action or to a submenu.

Graphical user interfaces are based on symbols, icons, colours or patterns that represent an action/process instead of text or descriptions of options. Selection is made by a graphic device such as a mouse. Graphic or text menu systems select an operation first followed by the object on which the action is to be carried out.

In batch mode searching, the system allows to store user profiles i.e. search strategies stored in the form of files and run them at regular intervals (e.g. SDI services)

9.5.6 Output/Report Generation Sub-System

The report generation sub-system allows the user to design the outputs in a desired format. The reports are generally executed on a final set. The module usually provides the user certain functions: to get records from the database based on the record reference numbers in a set; to identify the fields, their occurrences, and the sub-fields; to reformat data from system format to user required format (e.g. a date stored as 19970101 in the system may be output as 1-1-1997 or January 1, 1997); and a set of functions to control the execution of the report. One or more types of output formats are designed and the end user is usually provided with a set of output options (e.g. Citation format, full record with abstract, tagged format, formats suitable for export to other systems etc.)

9.5.7 System Usage Monitoring Sub-System

This module records the number of times a user logged in to the system, the time spent by the user in each search session, the CPU time utilized, the search terms, commands and strategies used, the number of records retrieved and so on. A statistical analysis of these log files will help to assess the usage of the system and the user interests and the gaps in the system also. For example, if the system is searched by a number of users for a a topic 'biochemical engineering' and the retrieved items are minimal, it means that the system is poor in providing information on the topic and additional efforts are required to fill up this gap.

9.6 FIVE LAWS *Vis-a-vis* ISAR SYSTEMS

The implications of Five Laws of Library Science, propounded by Dr.S.R.Ranganathan (exposed in 1928 and published in 1931) could be studied with respect to the design and development of ISAR systems. The Five Laws are so fundamental and perennial, they hold good for all times - past, present and future times. To make them more useful and relevant to the information age, they have been restated as follows:

1) Information is for use (Books are for use)

2) Every information user his/her information (Every reader his/her book)

3) Every piece of information its user (Every book its reader)

4) Save the time of the information user (Save the time of the reader)

5) The universe of information is ever growing (Library is a growing organism)

These corollaries of the Five Laws could be studied with respect to the objectives of ISAR systems stated by Pao (see 9.3.1).

The First *Law Information is for use* implies the promotion of use of information through information resources, which have relevant informational content directly related to the defined scope of the subject field. This refers to the 'Informational Content' of Pao's objectives of an ISAR System. The major implications of this law are -

* Make information available and accessible to the users at faster and affordable costs
* No barriers should hinder the information access between the users and the system

The Second Law *Every information user his/her information* refers to the Pao's specific user groups. Therefore, the implications of this law have direct bearing on the users. Every information system should have well-defined user groups.

* Identify the problems and needs of the information users
* Facilitate information retrieval for the information users through various modern technological devices for faster and quick information
* Serve to meet the needs and demands of the users effectively

The Third Law *Every piece of information its user* refers to Pao's 'Utility' of information resources acquired by the system. To ensure utility, every possible effort should be made by the information specialists/ designers of information retrieval systems. The major implications are-

* Develop facilities and programmes for browsing and processing of information
* Identify , analyse and provide access to information content to the users
* Evaluate the information in terms of clientele needs

The Fourth Law *Save the time of the information user* refers to Pao's 'Economics' and 'Performance Criteria'. The major implications are -

* Designing the systems conducive to users' searches
* Compatibility of the system to end-user search process
* Provision for information specialists to play as intermediaries

The Fifth Law *The universe of information is ever growing* refers to the multi-dimensional and continuous growth of information/literature. This also refers to the complex nature of the users' needs as well as growing number of information systems. The implications are

* Develop suitable indexing techniques and search tools
* Methodologies for handling information, especially interdisciplinary/ multidisciplinary nature.

9.7 LET US SUM UP

Let us recapitulate what has been discussed so far in this unit.

* Computer-based information systems can be categorized into: a) Information Storage and Retrieval (ISAR) Systems, b)Database Management Systems (DBMS), c) Management Information Systems (MIS), d) Decision Support Systems (DSS), and e) Question-Answering Systems (QAS).

* Information retrieval is a process concerned with the selection, representation, storage, organization, and accessing of information items to meet the specific requirements of a user community.

* The essential skills and knowledge required by an information scientist are - 1) Knowledge of the subject they are handling, 2) Current developments and trends in subject literature, 3) Knowledge of the vocabulary control systems, 4) Users' needs and their information seeking behaviour, 5) Methodology of database creation and representation of information in them, 6) Methods of evaluating the ISAR systems.

* Information storage and retrieval systems may be grouped into three types based on the content of the system: a) Reference retrieval systems, b) Document retrieval systems, and c) Fact retrieval systems

* The components (modules) of ISAR system are : a document selection subsystem; a data input and validation subsystem; an indexing subsystem; a vocabulary control sub-system; a search sub-system for interactive and batch mode operations; a user interface that helps the novice user in searching; an output/report generation subsystem; and a system usage monitoring subsystem.

9.8 REFERENCES AND FURTHER READING

GUINCHAT, Claire and Michel Menou. *General introduction to the techniques of information and documentation work.* Paris: Unesco, 1983.

PAO, Miranda Lee. *Concepts of information retrieval systems.* Englewood, Colo.: Libraries Unlimited, 1989.

SALTON, Gerard and Michael J. McGill. *Introduction to modern information retrieval.* Auckland: McGraw-Hill International Book Co., 1983.

VICKERY, Brian C. and Alan Vickery. *Information science in theory and practice.* London: Butterworths, 1987.

9.9 MODEL EXAMINATION QUESTIONS

I. ESSAY QUESTIONS

1) List out the various types of information systems and discuss the need for designing an information storage and retrieval system.

2) Discuss the objectives and functions of an ISAR system.

3) Explain the various components of an ISAR system.

II. SHORT NOTES

a) Five Laws vis-à-vis Objectives of ISAR Systems

b) Reference Retrieval Systems

UNIT - 10 : FILE ORGANISATION IN *ISAR* SYSTEMS

Structure

10.0 AIMS AND OBJECTIVES

The unit aims to introduce you to the file and record organisatioonn in Information Storage and Retrieval (ISAR) systems.

After studying the unit you could be able to

- describe the record and file structures in an ISAR system
- compare the fixed length and variable length records
- describe the structure of a flat file
- explain the representation of data, indexing methods and tools, and searching in an ISAR system.

10.1 INTRODUCTION

Having studied the objectives, components and types of an Information Storage and Retrieval System, one has to be clear about its file and record structures/organisation. File organisation or file design refers to the complete details about the type of information to be stored in the file. As you are aware that the files are constructed by the records, the amount of data to be stored in each record, the record structure, the relationship between different data elements, the methods/sequence of storing records and their access to be decided at the stage of designing an ISAR system.

Generally, three issues are associated with the file organisation - physical, structural and logical aspects. The physical aspects deal with the physical media of the documents (print, magnetic, optical, etc) on which the data/information is recorded and stored. This is concerned with the hardware features. The structural features are related to the organisation of data elements and records in the files. This leads to the third aspect, i.e., logical file structure. The speed of retrieval depends on how one understands the logical structure of a file. The following sections deal with the file and record structures.

10.2 FILE AND RECORD STRUCTURES

A file can be defined as a collection of information stored in a computer and handled as one unit by giving a single name. The information stored in a file can be handled by the software as a single unit or a collection of logically related units called *records.* Files created by text processing systems like Wordstar or Wordperfect handle the files generated by them as single units. Unlike files generated by text processing systems, that deal with continuous text, files in information retrieval system are organized in terms of units called *records* or *items.* Each record describes a single occurrence of an entity and all records in a file are homogeneous i.e. follow the same record structure. For example, an information storage and retrieval system file that contains descriptions of documents may include fields: personal names, title, year of publication, publisher etc. Similarly an index file contains records with fields: the indexed term, the document reference numbers. Access to information in information storage and retrieval systems is always expressed in terms of records i.e. a specific record matching a user query.

10.2.1 Record Organization

The files in information storage and retrieval systems can be categorized into linear lists, ordered sequential files, and indexed files based on the record organization and their maintenance within the file. All the above types of file structures are explained in the context of library and information science.

I) Linear Lists

A linear list is the simplest structure and literally contains an unordered collection of items. When new items are to be added they are always appended at the end of the end of the file, and the existing sequence of the records is not altered. Deletions from the file also do not need any file re-arrangements, resulting in the elimination of file maintenance process.

To answer a query, the records in a linear system are examined one by one till the required item is located. It is totally impossible to predict the exact location of a specific item in a linear list. Linear lists are useful when the file contains only a few items. But as the file grows the search process is time consuming. To overcome this limitation, the concept of clustered files is introduced. In a clustered file system, records are grouped in to classes based on mostly searched field value. For example, if a linear list contains description of books and the searches are made frequently on personal names, multiple files may be created on each alphabet i.e. the first file may contain names starting with alphabet 'A', the second file with 'B' and so on. When a new record is to be added, it will be appended to the appropriate file. When a classified file is used for retrieval, it is no longer necessary to examine every file item; instead the search can be restricted to certain classes of items that appear to be close to the request. But when a query needs to locate an item on another field value like 'Publisher', all the items in all the files are to searched.

Author	Jones	Ash	Brown	Adams	Smith	Scott	David
Title	Managing Computer Software	A note on the quality of software	Error Modelling in software	The Research directions of software engineering	Performance measures in software considerations	Software requirements and design software	Applying Graph Theory to
Topic	Software Computer Management	Software Quality	Software Error Model	Research Software Engineering	Software Performance Measures	Software Requirement Design	Software Graph Theory
Location Number	1	2....	i	k.....	n-1	n	new item

Figure-1: Linear List

2) Ordered Sequential Files

Similar to clustered files, ordered sequential files depend on the assumption that certain fields of records have special importance for retrieval purposes. The name of the author may be the main criterion to locate a journal article. These fields of importance are called keys.

Unlike linear lists, the items in an ordered sequential file are organized based on the key values. For example, if personal name is considered the key, the file is organized alphabetically. When a new item is to be added to the file, room should be made to enter the new item at the appropriate position in the file.

Author	Adams	Ash	Brown	David	Jones	Scott	David
Title	The Research Directions of Software Engineering	A note on the quality of software	Error Modelling in software	Applying Graph Theory to Software	Managing Computer software	Software requirements and design considerations	Applying Graph Theory to software
Topic	Research Software Engineering	Software Quality	Software Error Model	Software Graph Theory	Software Computer Management	Software Requirement Design	Software Graph Theory
Location Number	1	2	i	new item	k	n-1	n

Figure-2: Sequential file ordered by author

To locate a specific item, the search starts at the beginning of the ordered sequential files and works toward the end, item by item, till the desired item is found or a specific item's field value is greater than the searched value. To increase the efficiency of the ordered sequential files, a method called binary is introduced. For example, if a file contains seven items with key values A-B-C-D-E-F-G and the search is made for the key value 'C', the key value for the middle item in the file is compared with the key value in the search request. If the key values match, the search process is abondoned and the middle record is retrieved. If the values do not match, then it is checked whether the key value in the request appears before or after the key value of the middle record. If the value appears before, the first half of the file is considered for binary search else the second half of the file is considered, excluding the other half from the search process. This process continues till the desired item is located.

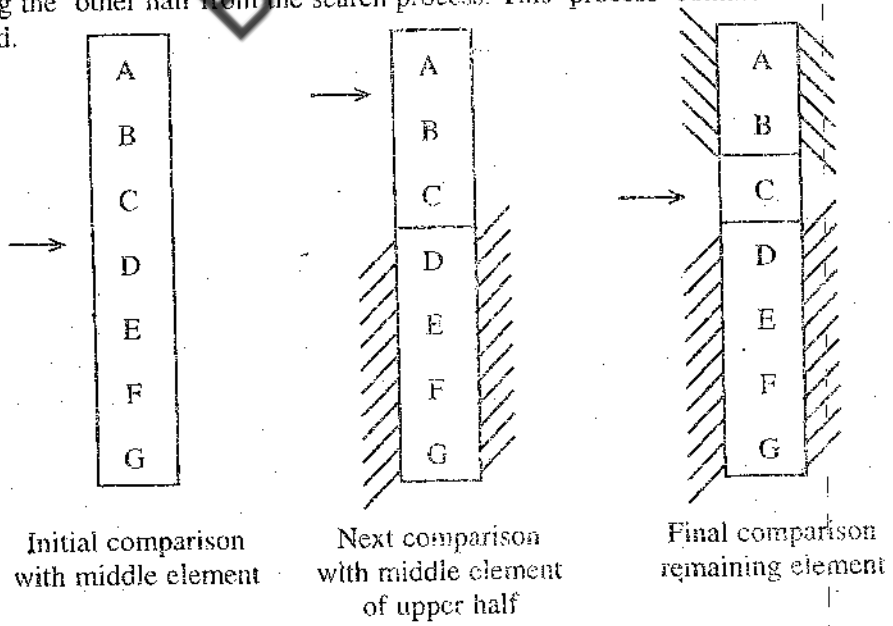


Figure-3 : Binary search example for search key 'C'

If a file contains 1023 items, a binary search method can locate the desired item in 10 steps on the average, while it will take 512 steps on an average to locate the same item in a linear list. Good examples of ordered sequential files are telephone directory, dictionaries.

Linear lists and ordered sequential files are also called direct files in which the items themselves provide the main order of the file and the searches are made directly on the document or record file, unlike indexed files.

3) Indexed Files

Indexed files are similar to linear lists in the context of record arrangement i.e. new records are always appended at the end of the file. But the search process does not depend on the sequential scanning of the file. One or more keys are identifiaed as searchable and indexes are created for these keys. Each record is identified by a document reference number or record number. The index file contains the key value arranged in order along with the document reference numbers associated with the key value. For example, an index file on personal names, may arrange the keys in alphabetical order, while another index file on year of publication may arrange the keys in numerical order. Unlike linear lists and ordered sequential files, indexed files provide access to records on multiple key values.

Related Information Items

Topic		
	Computer	1 3
	Information	1 2 4
	Retrieval	1 2 4 5
	Systems	1 3 4 5
	Users	2 5

(a)

Item Number	1	2	3	4	5
Author	Ash	Brown	Jones	Reynolds	Smith
Title	Aspects of Computerised Information Retrieval Systems	A survey of users of Information Retrieval Systems	The history of computer systems	The State of the Art of Information Retrieval	Users of new Retrieval Systems
Topic	Computer Information Retrieval Systems	Information Retrieval Users	Computer Systems	Information Retrieval Systems	Retrieval Systems Users

(b)

Figure-4 : Sample inverted file organization: (a) Index identifying record numbers corresponding to particular topics; (b) Sample information items

Search process is always carried first on the index file, the document reference numbers are retrieved and these document reference numbers are used to retrieve the records in the document file. Unlike ordered sequential files, the number of steps required to locate a desired item is reduced to the number of steps required to search the index for document reference numbers + additional steps required to locate the record in the document file.

When the document file grows, that is reflected on the size of the index files and these may become unmanageable. In such a situation, indexes are created to indexes and hierarchies of indexes are often required to minimize the search response time. The draw back will be an addition or deletion of an item may require changes to indexes at all levels.

The records in information storage and retrieval systems are of two types: fixed length records and variable length records.

10.2.2 Fixed Length Records

Each record in the document file occupies the same space and the fields within a record follow the same sequence. A single header includes the field definition i.e. the field names, their data types, and their sequence in the record. The fields whether provided with a value or not, occupy the same space in the record. For example, if field author is allotted 15 characters and the entered field value size is only 10 characters, the remaining 5 characters are padded with blank spaces, there by resulting in the fixed size record. Fixed length record definition allows the programmer to predict the position of a specific record and a field therein by simple computations.

author (15)	title (20)	year (4)
Oberoi, S.	Computer science	1996
Date, C.J.	Database systems	-

Figure-5: Fixed Length Record

In the above example, the second record starts at the 40th position and the title field starts at 55th position in the document file. This type of record definition is very much suitable when the fields are of fixed number and limited size and all fields are supposed to have a value entered for them. Good examples are directories and mailing lists. Design of systems on fixed length record type is predominant in RDBMS environment. The file is considered to be a two dimensional table (flat file) where the rows designate a record and each column a field. Repeatable fields are not allowed and each field should be atomic in the sense that it should carry only one value. In this sense the records are also considered to be flat.

Author 3	
Author 2	
Author Title year	Author 1 Title year

Figure-6: Flat record Non-flat record

The field sequence is critical to data processing in this environment. The single header is read first before switching to records and their field values.

10.2.3 Characteristics of Bibliographic Data

The applicability of fixed length record types for library and information applications is to be considered now. The information stored about documents is basically of narrative type and expressed in natural language as in title and abstract fields. The data is basically static and cumulative in nature. Document descriptions in a library environment express specific characteristics: one more occurrences of fields (repeatable fields) i.e. a document may have no author, one author or multiple authors; variable length fields (size of the title of a document may vary from record to record; variable or inconsistent occurrence of field values (edition statement may be required if 2nd or above edition; conference details are included only when the document belongs to conference proceedings etc.); fields may have subfields (e.g. imprint field include place, publisher and date of publication as subfields). If maximum space is allocated for all possible repeatable fields, there will be an overhead of disk space and results in underutilization of storage resources.

The above problem can be solved by two ways: creating a document file that includes mandatory fields and supporting files with optional fields. For example, a main document file that includes first author, title, year of publication, pagination, publisher etc. may be created. The supporting file(s) may include fields like joint authors, edition, conference name, date, and place, etc. Every document will have a record in the main file. A record in the supporting file is created when the document has certain optional data elements to be included. This type of organization results in the distribution of data about a single item among multiple files and a query needs to search all these files to extract complete information about a single item. This organization increases the query response time as multiple files along with their indexes are to be opened and searched.

To meet the specific requirements of bibliographic data descriptions, two concepts are introduced: variable length records and flat files.

10.2.4 Variable Length Records

Each variable length record may contain one or more fields and the field lengths may vary from record to record. Fields with no value are totally excluded from the record. Repeatable fields are allowed and each occurrence of a repeatable field is identified by a serial number like first occurrence of author (author 1), 2nd occurrence of author (2nd author). Unlike fixed length records, a comprehensive data definition or header includes the details of all fields, their types, and their repeatability characteristics in the form of a table. A specific header for each record is used to identify the fields included in that record, their sequence and data lengths. and the total record length. The international standard ISO 2709 (described in detail in other chapters) is a good example of variable length record representation in information storage and retrieval systems. In this standard, only numbers are allowed as field identifiers. As per this standard, each specific header for a record has three parts: a fixed length (24 characters) label, a variable length directory and a variable length data. The label includes information on the total size of the record, the status of a record (new or modified etc.), the length of field identifiers, length of subfield identifiers, base address of data etc.. The directory includes an entry for each field that includes the field number, length of data field and starting character position. If there are 10 fields in a record there will be 10 directory entries. The directory is followed by actual data values and each field is terminated by a field.separator. Fixed sequence need not be followed as specific header will take care of it. As each record is of different size, and as it is not possible to calculate the position of a specific record, a separate file is maintained to store the relative positions of each record in the document file.

Kernighan, B.W	Ritchie, D.M.	Elements of C 1988
Oberoi, S.	Computer science	1996
Date, C.J.	Introduction to database systems	2nd edition

Figure-7: Variable Length Records

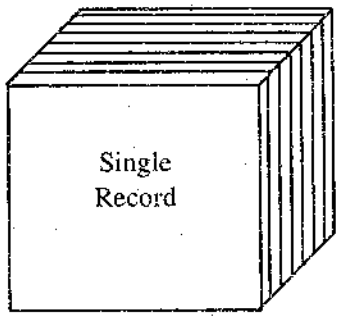
10.2.5 Fixed *Vs.* Variable Length Records

A comparative account of the fixed length and variable length records will provide a better understanding of their structure and organisation.

	FIXED LENGTH RECORDS		VARIABLE LENGTH RECORDS
1)	Fixed length fields	1)	Variable length fields
2)	Fixed sequence of fields	2)	Field sequence may vary from record to record.
3)	Fields with no value are filled with blank spaces. Fields with less size are padded with blanks.	3)	Fields with no value are excluded from the records
4)	Leads to overhead of disc space	4)	Leads to disc space conservation
5)	A single header includes field names, their types, sizes and sequence. Not following the sequence leads to misinterpretation of data.	5)	A general header includes the details of field names, their types, possible sizes, repeatability, etc. A specific header for each record contains the details of fields included, their sequence, size, their relative position within the record, etc.
6)	The position of a record and the fields there in can be calculated	6)	A separate file is required to maintain relative position of a variable length record
7)	Modification of a record will not change the record position as the modified record is record is overwritten on the existing record	7)	Modified record is always appended. Consequently the cross reference file to be updated for every modification
8)	Access to specific record and the fields therein is fast as the positions can be easily calculated.	8)	Access to records is through cross reference file only. Once record position is acquired the specific record header is to be read for field values. This is relatively slow.

10.3 FLAT FILE SYSTEM

The term Flat file is also used in another context, where the software allows a designer to define a single record type for the entire system (e.g. CDS/ISIS).



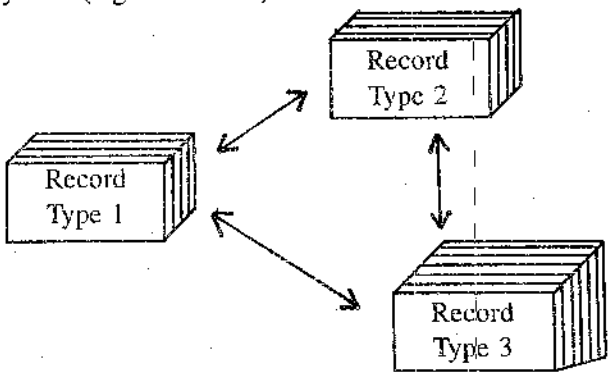


Figure-8: Flat file Non-flat file

When two or more types of document/record descriptions are to be included, the designer has to create a comprehensive data structure that includes all the fields/data elements of the document record types. For example, if the system includes descriptions of research projects and documents generated by those projects, the data structure should include all the data elements pertaining to both descriptions. Flat file systems adopt the variable length record concept and are best suited for bibliographic descriptions.

10.3.1 Usage

A flat file system is preferred when:

* data included is static in nature i.e., changes to data values are made very rarely.
* data is cumulative i.e. new records always added to the file and deletions are minimal.
* Each record is of variable length and repeatable fields are to be accommodated
* Data related to a single item is to be kept at one place i.e., in a single record as fast retrieval predominates redundancy control or disc space conservation aspects.
* Mostly narrative or text oriented data are included.
* The system is not transaction oriented.

These systems usually adopt batch update method that will not affect regular services. E.g.: Bibliographic databases.

10.3.2 Characteristics

The major characteristics of the flat file systems are -

1) A single record type only can be defined. To accommodate multiple record definitions, a comprehensive data structure that includes all the fields/data elements of all the record types is to be defined.

2) Each record should be comprehensive i.e., data related to an item should be with in a record.

3) No sharing of data among records or database files

4) Each record is identified by an artificial key either assigned by the software or by the user

5) Access to data is through the unique key

6) Searches are through secondary key indexes that point to records by unique keys

7) At a time only a single database/document file can be opened along with necessary indexes.

8) Software does not provide facilities to extract data from two or more databases for a single operation

9) Inverted files are not updated automatically. A separate utility is to be run to update the indexes. While updating, the database is usually locked. Instant update facilities are not normally provided.

10) Deferred/batch updating does not allow to answer instant queries

10.3.3 Data Definition

Data definition details are not tightly coupled with the database file. Details are maintained as a separate table file, called Field Definition Table.

* Each field is identified by a tag (numerical/ alphabetical) and its characteristics are defined in the field definition table/DDL file

* Addition of a field will not be reflected in old records. If necessary each record is to be accessed and supplemented with the field value.

* Addition/deletion of a field would not demand for database file reorganization. The necessary modification is carried out in the comprehensive data definition table only. The specific header will take care of modifications at record level.

* Deletion of a field from the field definition table will not be reflected in the old records i.e. entire file need not be reorganized. But the specific field is not accessible even the old records have included values for the field till the field definition table is updated.

10.3.4 Record Structure

Each record will carry three parts: label, directory and data.

Label - includes information on the record size, tag size, subfield identifier length, base address of data etc.

Directory - includes an entry for each field with field tag, field length and address details. No entries are made for missing fields. Additional entries are made for repeatable fields

Data includes actual values of fields and each field value is terminated by a field separator.

10.3.5 Data Manipulation

Flat file systems allow the following types of data manipulation:

Addition- New record is always appended at the end of the database file

Deletion- An existing record is marked for deletion but not physically removed

Modification - Old version of the record is marked for deletion. Modified record is always appended.

10.3.6 Querying

The Querying in flat file systems include :

Multiple key searches with Boolean operator application is supported.

Retrieval is always in terms of sets (not in tables) that contain unique key values assigned to record, which satisfy a condition.

Refinement of a search is carried out by combining sets or embedding a set number as a parameter in the query.

As sets contain the document reference numbers/pointers to records (not the values of fields of records), the entire record is available to the user for any manipulation.

10.3.7 Data Privacy and Security

Data security is provided through backup and recovery utilities usually provided with the software. Some software like BASIS and BASIS+ provide before and after journaling facilities i.e. storing in a separate file the before and after images of the records modified, particularly in an interactive data entry or update session. In case of system failure, these images/ versions of the added or modified records can be played back to the data file.

Privacy is provided through user accounts and passwords; Assigning privacy codes to records at record or field level and privacy codes to users also. For example, if a user is assigned

a privacy code of '4' and the condition to access the records is ' = ', then the user is allowed only the records with the privacy code '4'. Similarly, other relational operators " < <= > >= != " can also be used as conditions.

10.3.8 File Organization

In flat file systems, when the user deletes a record, it is not physically removed from the file but only marked for deletion. Similarly, when a record is modified, the older version of the record is marked for deletion and the new record is always appended. If a record is modified three times, all the 4 versions of the record will remain in the file. So if a file has 100 records and all the records are modified at least once, the date file size will be doubled as it contains both the versions of the records. This growing size of the data file due to modifications may lead to slow response while searching and underutilization of disc space.

To overcome these problems, the file is to be reorganized so that it retains only the current versions of the active records. The older versions and deleted records are to be physically

removed from the files, the associated indexes are to be updated and the disc space that released is to be returned to the operating system. Some software take care of these processes automatically, thereby relieving the database administrator from frequent file organization processes (e.g. BASIS, BASIS +). In other software, separate utilities are provided for file reorganization. The database administrator has to lock the database, run these utilities at regular intervals (e.g. Backup and recovery procedures in CDS/ISIS).

Differences in file and record organization play a crucial role in maintenance and retrieval processes. The software available in the market adopted one or more concepts discussed above to meet the specific requirements of their customers. For example the well acclaimed Unesco product CDS/ISIS adopted the flat file (single and comprehensive record) model specially meant for narrative (textual) and static type data; ISO 2709 format as basis for defining internal structure with variable length record descriptions and batch mode processing. The software also includes record locking facility in the multi-user environment. Some DBMS applications for integrated library automation are designed and developed using this software (e.g. SANJAY) but this required extensive programming.

Similarly, another software BASIS Plus, marketed by NIC (National Informatics Centre) in India adopted a non-flat (multiple record definitions) file model allowing variable length record descriptions. It allows to extract data from one or more record types by joining them on a common key value (a feature in RDBMS systems). It adopted a record structure similar to ISO 2709. It provides instant and batch updating facilities and the designer can select appropriate one based on the time taken for updating and the system requirements for each record type. For example, document descriptions can be kept for deferred or batch updating as it is time consuming when a number of indexes are to be updated. On the contrary, transaction files like records in a circulation system can be updated immediately as these subsystems have to meet instant queries. All the concepts of RDBMS (data redundancy control, concurrency control, integrity checks etc.) are taken care while designing the system. The software also comes with an integrated library automation package called TECHLIB- PLUS that includes Acquisition, cataloging, OPAC, circulation and serials control modules.

10.4 FUNCTIONAL APPROACH TO INFORMATION RETRIEVAL

Every information storage and retrieval system can be described as consisting of a set of information items i.e. documents, a set of requests, and some mechanism to match these two and retrieve desired items. As mentioned earlier, the information storage and retrieval systems basically deal with narrative type data i.e. in natural language. User requests are also received in narrative type.

e.g. The following two titles deal with information science:

1. *Information science in theory and practice*
2. *Introduction to modern information retrieval*

A user requests documents on information storage and the second user on information systems. In this case the relevance of the two documents to the requests cannot be determined directly as the titles are not carrying the terms, 'information storage' 'or 'information systems '. In a manual search of the library catalog, a library staff may assist the user to locate these two documents indexed under the term 'information storage and retrieval systems'. A similar mechanism is required to match the user requests with the document details in a computer-based retrieval system.

10.4.1 Data Representation in ISAR Systems

To get over the above problem, the documents or information items are converted to a special form using a classification or indexing language and this process is termed as mapping In the above example the two documents may be indexed under the term 'information storage and retrieval systems'. Similarly the user requests are also converted into a representation consisting of elements from the indexing language. Thus the two requests above can be represented by the term 'information storage and retrieval systems'. A search on the system matches the user request expressed in the index language terms with the items and retrieves both the titles as there is a perfect match.

The mapping process of documents and requests may be carried out manually, automatically or a by a combination of the two processes. For example, in an information storage and retrieval system which uses an on-line thesaurus as an indexing aid, the indexer may select the preferred terms manually. The validation procedures provided with the software reject the records with terms which are not included in the thesaurus. When a non-preferred term is included in the record, the software automatically indexes the record under the preferred term. Similarly, a switching from non-preferred to preferred term will take place during the user query process. To get the precise items from the system, both the mappings should be carried out using the same indexing language tool. It is also to be noted that a user query may not be restricted to concept or subject based retrieval. To meet the other requirements such as search by personal names, corporate names, date of publication etc. additional indexes are to be created. All this means that information storage and retrieval systems depend heavily on indexed files.

10.4.2 Indexing Methods and Tools

The American Heritage Dictionary defines an index as " an alphabetized listing of names, places, and subjects included in a printed work [which] gives for each item the page(s) on which it may be found. As defined by American National Standard Institute (ANSI) " An index is a systematic guide to items contained in, or concepts derived from, a collection. These items or derived concepts are represented by entries arranged in a known or stated searchable order, such as alphabetical, chronological or numerical". The above definitions may be extended to include indexes created to databases.

The purposes of an index are:

* to facilitate reference to the specific item;

* to give nomenclature guidance through see and see also references.

The principal component of an index is the entry, which consists of a heading, most commonly a keyword or phrase used to identify a subject, and a locator i.e. page number(s) in a book index, a record number in an information storage and retrieval system.

Indexes can be categorized as pre- and post-coordinated indexes. In pre-coordinated indexes (e.g. the usual non-manipulative, published indexes and subject catalog cards), the correlation of the words in the headings and modifications provides the required selectivity. for example, the term 'Automation' is broad enough to index banking systems, libraries, machine tools and so on. The index entry 'Automation, of library circulation systems' is much more specific than the heading 'automation'. Since correlations are made during the indexing process and prior to the use of the index, it is called a precorrelative/precoordinated index. The user has to follow the specific sequence of terms used to locate a record of his interest.

In post-coordinated indexes (e.g. computerized indexes), the correlation takes place at the time of search. The index terms are not arranged in a predefined manner. For the above example, three terms can be selected : Automation, libraries and circulation systems. The record is indexed under these three terms and a correlation of these terms with a Boolean operator AND will retrieve the desired record.

The indexes created in an information storage and retrieval system can be categorized into alphabetical indexes, chronological and numerical indexes. Subject, author, title indexes come under first category. The possible searches that can be made on alphabetical indexes are exact matching, left or right truncation and wild card searching. The relational operators like '> >= < <= = !=" can not be used on these indexes.

The other two kinds of indexes, chronological and numerical, allow the application of relational operators while searching. For example, a query can include a condition to retrieve all the documents published after a specific year expressed as '> 1996'. Similarly, application of Boolean operator AND will help to retrieve documents published between two years (>= 1995 AND <= 1997).

1) *Manual Indexing*

As mentioned earlier, subject searching predominates in information storage and retrieval systems. Subject indexing may be manual or automated. In a manual system, the indexer analyzes the subject content of the document by going through the title, contents pages, preface etc. After assessing the the content, he selects specific words or phrases that can represent the subject fully. Then he uses a standard controlled indexing language tool like thesaurus to select lead terms/acceptable terms. As the indexer analyses the subject content of a document and assigns subject terms, it is called assigned indexing. The manual indexing methods demands for a better understanding by the indexer of the document collection characteristics and the type of user queries the system may be expected to process in the future. Further more, all the indexers should be consistent while indexing to guarantee that similar documents are identified by comparable indexing entries.

It may also not be possible for the indexers to predict the future requirements of a user community. For example, a concept considered to be of no importance and excluded by the indexers during subject analysis of a document, may become primary in future when the policy of the organization that the information storage and retrieval system serves, changes. To answer the queries on this new concept, the entire collection is to be reindexed. The other solution will be to provide an exhaustive index for all the mandatory and non-mandatory concepts covered in the documents. This is not only time consuming on the indexer's part, but the growing size of the index file will also decrease the system performance.

2) *Automated Indexing*

Automated indexes are based on the assumption that ideas are communicated by words and by their arrangement (context) and 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. For computational purposes, a word is defined as a sequence of symbols, either alphabetic or numerical, separated by spaces. In automated indexing systems each significant word is indexed excluding articles, conjunctions, prepositions and certain common and non-significant words like "always, there, here, therefore, where, when, thereby" etc. which are usually stored in a file called 'stop word file'. If a field content (e.g. abstract) is selected for indexing, the software first creates a list of terms from the abstract excluding the stop words; makes index entries that include locator details of each word: the record number, field number, paragraph number, sentence number, and the word number with in each sentence. If the abstract field (e.g. Field 100) of a document number 345 has the phrase 'information retrieval' in the second sentence as 4th and 5th terms, the possible index entries will be:

information	345	100	1	2	4
retrieval	345	100	1	2	5

The document 345 could be retrieved by the query statement.

Information ADJ retrieval

where the operator ADJ indicates that the two terms should appear together with in a

sentence. As word position information is available, one can also insist on a positive difference between location numbers of words. If the difference is allowed to be either positive or negative, the order of the terms is disregarded. In DIALOG system, the order of the words is important and the systems assumes they are to appear in the order specified in the query statement. Thus, Select Information (5w) retrieval would find all documents in which the term 'retrieval' follows the term 'information' with in a distance of up to t words.

As all the significant words are indexed along with their location details, it is possible to count the number of times a given word appears in a document and then judge the relevance of the document to the query. Truncation makes it possible to count all appearance of the same stem as one word type. For example, software can combine the counts of the stem 'index' with the counts of 'indexes, indexer, indexers, and indexing'. Software can also count the number of times that word indexing is preceded by the term 'automatic', and select the documents that contain these two words and present them in the order of frequency. An on-line tool like thesaurus can also be used to include synonymous, narrower terms, and spelling variations also while counting the frequency of words.

e.g. Draught (British form)

Draft (American form)

While counting the frequency of either of the terms, both the British and American forms may be taken into consideration.

Automated indexing methods help to answer the unpredictable queries. Most of the information storage and retrieval systems adopt both the methods : a limited number of subject terms are assigned to each document by a manual subject analysis; and supplemented by an automated word index. In both the cases, the support of an on-line thesaurus is fully recognized.

10.4.3 Searching on ISAR Systems

Each information storage and retrieval system consists of a document file and one or more auxiliary files known as inverted files. Each record in the document file is identified with a unique record number known as document reference number. The inverted files contain the indexing terms and each term is associated with a list of document reference numbers also known as postings. The document reference number uniquely identifies a document to which the index term has been assigned.

A search on the system includes two processes: to check the index file for the desired term and retrieve the associated document reference numbers and to use these reference numbers to retrieve the records from the document file. Once the index file is searched for a specific term, the associated document reference numbers are saved in to a set. A set is nothing but a table that contains the document reference numbers and uniquely identified by a serial number for the search session.

An entry in the index file may look like this:

Term	Document reference numbers							
Library automation	1	20	35	37	63	64	91	101
Information retrieval	1	20	35	64				
Information services	1	25	37	63	78	91		

When a query is made on the term 'information retrieval' in a new search session, the four document reference numbers are retrieved and saved as set #1 with the label 'information retrieval' for future identification of the set. When a second query is made on the term 'library automation', the retrieved document reference numbers are saved as set #2 with appropriate label.

set no	label	no of records
Set #1	Information retrieval	4
Set #2	Library automation	8

The document reference numbers are used to retrieve the associated records from the document file either for display on the terminal or for a report generation. The sets are available to the searcher till the end of the search session.

Queries consisting of a single term are rarely used. The mapping process of the complex user query may result in two or more terms or a concept dealt within a document may be represented in two or more terms in the system using a specific indexing language. For example, the concept "library automation" may be represented by two terms " libraries" and "automation" in a system. Searching the system by one of the terms only may result in noise i.e. unrelated document references. Similar is the case with the user query on documents on 'information systems and services' which is represented by the terms `information storage and retrieval systems' and 'information services. If no mechanism is provided to make a combined search by using two or more terms, the user has to scan the documents retrieved by using a single term, one by one, for the presence of other terms also to get a precise list of documents and it is definitely a time consuming process.

1) Boolean Searching and Boolean Operations

To identify the documents that contain both the terms ' libraries' and 'automation', it is necessary to process the information retrieved from the index files rather than the information from the document file. Boolean logic is used to construct the queries consisting of a variety of terms using the Boolean operators AND, OR, and NOT. These operations are implemented by using set intersection, set union, and set difference procedures, respectively.

The query which may be used to identify the documents on 'library automation' may be stated as

libraries AND automation

The steps involved in finding the document include:

1) Use the inverted file to retrieve the document reference numbers associated with the term libraries and save those references in set #1

2) Use the inverted file to retrieve the document reference numbers associated with the term 'automation' and save those references in Set #2

3) Determine the document reference numbers that constitute the intersection of the sets 1 and 2 i.e. the documents reference numbers that are contained in both the sets. Save the set as set #3.

4) Use the main document file to retrieve the document identified by the document reference numbers in set #3.

set no	label	no. of references	document reference numbers
Set #1	Libraries	4	1 3 11 25
Set #2	Automation	3	3 25 27
Set.#3	Set #1 AND #2	2	3 25

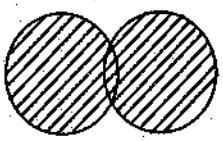
A diagrammatic representation of the above query process may look like:

Figure 9: 1 and 2

Similarly, the query 'libraries or information centers' refers to documents which are identified either by the term 'libraries' or by the term' information centers' or by both the terms. Set #1 and #2 are determined in the manner as for the 'AND' operator. These sets are then combined into a new set #3 which includes document reference numbers contained either in Set #1 or #2 or both set by the process union.

set no	label	no. of references	document reference numbers
Set #1	Libraries	4	1 3 10 25
Set #2	Information centers	5	1 10 29 31 42
Set #3	Set #1 OR #2	7	1 3 11 25 29 31 42

The diagrammatic representation of a union of two sets looks like



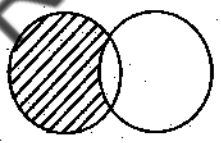
1 or 2

Figure:

A query on 'information services in countries other than India' refers to documents identified by the term information services but not dealing with the Indian context. The sets #1 and #2 are determined in the same manner above but the third set with be the result of set difference procedure i.e. the document reference numbers of set #2 are excluded from set 1 while determining set #3.

set no	label	no. of references	document reference numbers
Set #1	Information services	5	13 17 19 28 45
Set #2	India	3	17 28 49
Set #3 Set #1 NOT #2		3	13 19 45

The diagrammatic representation of a set difference procedure looks like



1 NOT 2

Figure 11:

2) *Order of Operations*

When a query includes more than one operator, the complexity of the query grows substantially and a variety of rules are necessary to ensure that queries submitted are interpreted correctly by the system. For example a query in natural language read like this:

Information services in India or Pakistan

and the query formulation could be

Information services AND India OR Pakistan

If the processing starts from the left and works toward the right the result will be

Set #1 Information science
Set #2 India
Set #3 Set #1 AND #2
Set #4 Pakistan
Set #5 Set #3 or #4

where Set #5 includes documents consisting all the documents on Pakistan on all subjects.

If the process starts from the right and works toward the left the result will be

Set #1 Pakistan
Set #2 India
Set #3 Set #1 OR #2
Set #4 Information services
Set #5 Set #3 AND #4

and this is right interpretation of the user query. Different information storage and retrieval systems follow different rules and it is necessary to strictly adhere to them to avoid noise in the search output. For instance, one procedure specifies that:

all the OR operators are performed;
then AND operators;
finally NOT operators;
all equivalent operators are performed from left to right;
operations in parentheses are performed first.

In this case, to avoid wrong interpretation of query, parentheses are usually provided to circumvent the strict processing order described above. Under these rules of precedence of operation, the above query may be formulated as:

Information services AND (India OR Pakistan)

Similarly, the query ((information services OR information systems) AND (Asia OR Europe)) NOT India is processed as:

Set #1 Information services
Set #2 information systems
Set #3 Set #1 OR #2
Set #4 Asia
Set #5 Europe
Set #6 Set #4 OR #5
Set #7 Set #3 AND #6
Set #8 India
Set #9 Set #7 NOT #8

Set numbers may also be included as parameters to another query. e.g. Set #5 and automation. Similarly, sets created by searching other fields like author, title, publication date, ISBN numbers etc. may be combined with sets created by subject searches for refinement purposes.

Searches also may be made using truncated terms (left or right). For example a search on the truncated term 'index*' retrieves all the records that carry a term with the prefix 'index*' like 'index, indexes, indexing, indexer, indexers' etc. Similarly the term '*gonism' with left truncation will retrieve all the records that include a term with the suffix 'gonism'. A search term expressed as 'An*gonism' (termed as wild card searching) will retrieve all the records with terms that have a prefix 'An' and a suffix 'gonism'

What ever be the search strategy i.e. limited to subject search only, or combined with other field searches, the retrieval is only in terms of sets as far as information storage and retrieval systems are concerned.

10.5 LET US SUM UP

Let us recapitulate what has been discussed so far in this unit.

* A file is a collection of related data/information stored in a computer and handled as one unit by giving a single name.

* Files in ISAR systems are categorised into linear files, ordered sequential files, and indexed files based on the record organisation and their maintenance within a file.

* The records in an ISAR system are of two types: fixed length and variable length.

* The software (e.g.: CDS/ISIS) allows a designer to define a single record type for the entire system. This is called 'Flat file system'.

* The indexes created in an ISAR system can be categorised into alphabetical, chronological and numerical indexes.

* In automated indexing systems, the subject of a document can be derived by a mechanical analysis of the words and by their arrangement in the text.

10.6 REFERENCES AND FURTHER READING

GUINCHAT, Claire and Michel Menou. *General introduction to the techniques of information and documentation work.* Paris: Unesco, 1983.

SALTON, Gerard and Michael J. McGill. *Introduction to modern information retrieval.* Auckland: McGraw-Hill International Book Co., 1983.

VICKERY, Brian C. and Alan Vickery. *Information science in theory and practice.* London: Butterworths, 1987.

10.7 ASSIGNMENT

Critically examine the record and file structures of any library application software you have been using in your library and information centre for information retrieval.

10.8 MODEL EXAMINATION QUESTIONS

I. ESSAY QUESTIONS

1) Explain the file and record structures used in ISAR systems.

2) Bringout the differences between fixed length and variable length records.

3) Explain the functional approach to information retrieval.

II. SHORT NOTES

a) Boolean Operators

b) Flat files

UNIT - 11 : EVALUATION OF *ISAR* SYSTEMS — METHODOLOGY

Structure

11.0 AIMS AND OBJECTIVES

Evaluation of an information storage and retrieval system means to make a judgement about its worth or merit. This unit aims to explain you the process of evaluating an ISAR System.

After studying this unit, you should be able to

- list out the stages in evaluation
- explain the various measures of system effectiveness
- discuss the failure analysis and the methods of improving the system performance
- explain the concept of cost-effectiveness and cost-benefits.

11.1 INTRODUCTION

Evaluation of information storage and retrieval systems is the process whereby various parameters are used to measure the performance of the existing system, the causes for the mis- or not up-to-the mark behavior of the system are identified and a set of remedial measures are suggested to improve its performance. The first step i.e. performance testing is termed as evaluation of effectiveness, followed by the second as diagnosis or failure analysis and the third as therapeutic study.

The first step involves measuring and expressing the performance of the system as a whole according to some type of quantitative scale and known as *macro-evaluation.* The next two steps involve a detailed study of the subsystems of an information storage and retrieval system (e.g.: indexing policy, indexing language, indexing methods, searching methods, coverage etc.) and is known as *micro-evaluation.* Evaluation of an information storage and retrieval system may be carried out at any of several stages of system development: the experimental or conceptual stage, the prototype stage or the fully operational stage. It should be clearly noted that evaluation of a system is not carried out as an intellectual exercise but is supposed to lead to significant improvement in the performance of the system.

Evaluation of an information storage and retrieval system could be completely subjective i.e. end-users could be asked to assess the value of the services provided by the system, usually through a questionnaire. Subjective evaluation based on filled questionnaires gives some idea of user satisfaction but could not be diagnostic in the sense it could neither find out the causes for the poor performance or failure of the system in question nor could lead to the improvement of the system.

Objective or quantitative evaluation makes use of some type of performance figures to express the degree of success of a search or service provided by the system. These performance figures are helpful in differentiating poor searches or services from the good searches, which are further, subjected to detailed analysis to determine the causes of failure.

11.2 STAGES IN EVALUATION

As mentioned earlier, evaluation of an information storage and retrieval system may be carried out at two levels: macro-evaluation or micro-evaluation. The evaluation may be limited to a subsystem like indexing language and would like to assess the comprehensiveness of the language to represent various concepts of the subject areas covered in the system, the

degree to which relationships among these terms are established (e.g. in an alphabetical listing of subject terms, relationships of terms such as synonymous, broad, narrower or related terms are not established but in a thesaurus these are established), the currency of terms used in the indexing language and the like. Similarly, a user-interface may be evaluated to assess the support it extends to a novice user.

Evaluation program comprises a number of stages:

* Establishing the scope and purpose of the programme (i.e. deciding what to evaluate)
* Designing the evaluation
* Conducting the study and gathering data
* Analyzing and interpreting results in terms of failure or success
* Identifying the causes for poor performance of the system at each subsystem level
* Suggesting remedial measures to improve the performance of the system at different subsystem levels
* Making system modifications designed to improve the over-all performance level
* Evaluating the modified system to assess the impact of modifications carried out on the system

It should be noted that although an evaluation exercise may concentrate upon one particular subsystem (e.g. indexing policy and procedures), it cannot be evaluated in isolation, as the various subsystems (indexing, indexing language, searching and user-interfaces) are closely interdependent. A change or modification in one subsystem will have effects and repercussions elsewhere. For example, the indexing policy may direct the indexer to user narrowed or stringent terms to represent the concepts covered in a document but the indexing language/ controlled vocabulary may not be comprehensive and including these terms. Unless the indexing language subsystem is updated, the indexing policy could not be successfully implemented. E.g.: A thesaurus entry under the term 'allergies' may look like:

Allergies

arthropod allergies

* drug allergies * omitted terms in a thesaurus

farmer's lung

food allergies

* milk allergy

humidifier disease

For example, a document dealing with drug allergies is to be indexed. The indexing language contains only a higher-level generic term 'allergies' or the term 'drug allergies' is omitted at a lower level. In this case the indexer is forced to select the higher level term only. Similar situation may occur at search level also. If a document on 'milk allergy' is indexed on a higher-level generic term 'food allergies', the document can be retrieved by using the term 'food allergies' only.

It should be noted that compromising with low quality at one subsystem level might effect the overall performance of the system.

11.3 SCOPE OF EVALUATION

Evaluation of an information storage and retrieval system could be in terms of:

* system effectiveness - evaluation of system performance in terms of degree to which it meets user requirements

* cost-effectiveness evaluation in terms of how to satisfy user requirements in the most efficient and economical fashion

* cost-benefits evaluation of the worth of the system itself i.e. the overall benefits the organization achieved in terms of improved decision making, avoidance of duplication of research, avoidance of loss of productivity etc.

While evaluation of system effectiveness is of direct concern to the users of the system, cost-effectiveness evaluation is of concern to system operators and managers, cost-benefits evaluation is of concern to top management.

11.4 EVALUATION OF SYSTEM EFFECTIVENESS

As mentioned earlier effectiveness evaluation concentrates on the user requirements. The fundamental requirements of users could be:

* Coverage of the information retrieval system in terms of comprehensiveness and currency

* Recall

* Precision

* System response time in answering a query

* User effort involved to attain high recall and precision ratios in search output

* Form of output i.e. the degree of description of documents/document surrogates presented to the user

11.4.1 Coverage

Coverage of an information storage and retrieval system could be expressed in terms of:

* the types of documents included (monographs, chapters in monographs, journal articles, reviews, patents, theses, unpublished documents etc.)
* the period of coverage
* the amount of description of these macro- or micro- documents (e.g. descriptions with or without abstracts)
* the time taken for a document to find its place in the system i.e. the time lag between the actual publication date and the entry date in the system
* the percent of documents included in the system i.e. whether all the literature published or unpublished included
* the percent of documents included in the system on the topics of interest to the user

11.4.2 Recall

User approaches a system with a view that he can retrieve one or more relevant documents that could satisfy his information requirements. The degree of a system's success in retrieving relevant documents from its data source could be expressed quantitatively as

$$\frac{\text{the number of relevant documents retrieved by the system}}{\text{the total number of relevant documents contained in the system}} \times 100$$

Suppose there are 10 relevant documents on a particular subject in the system and a subject search conducted using normal procedures retrieves 7 of these 10 documents, the recall ratio is (7/10) x 100 or 70%.

Recall ratio is also called *hit rate, sensitivity,* and *conditional probability of a hit.* The methods of estimating relevant documents contained in a system are discussed in later sections.

100% recall ratio could be achieved by broadening the search strategy and retrieving a large portion of the collection. Information storage and retrieval system serves as a filter where it lets through what is wanted (recall) while holding back what is not wanted (precision).

11.4.3 Precision

Precision ratio, also referred to as *relevance ratio,* may be defined as:

$$\frac{\text{the number of relevant documents retrieved by the system}}{\text{the total number of documents retrieved by the system}} \times 100$$

Suppose against a search strategy, there are 50 documents retrieved by the system and out of them 10 documents are judged as relevant by the user, then the precision ratio will be

(10/50) x 100 or 20%

The two measures recall ratio and precision ratio jointly indicate the filtering capacity of the system. Achievement of 70% recall ratio at a precision of 50% indicates greater efficiency of the system than the attainment of the same 70% recall ratio at a lower precision, say 20%.

Recall and precision ratios are inversely proportional i.e. the efforts to improve recall by broadening the search will tend to reduce precision, while the efforts to improve precision narrowing the search will tend to reduce recall, as expressed in Figure 1.

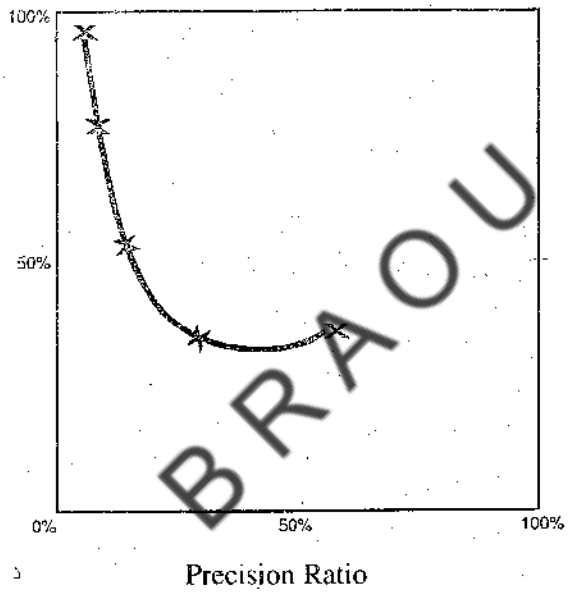


Figure-1: Performance curve of recall versus precision

Information retrieval systems are usually designed to provide higher recall and precision ratios. But there will be occasions where a search strategy may be confined to higher recall ratio only, while the other on higher precision ratio. For example, a user who is writing a book or review article on a particular topic wishes to see everything included in the database and his requirements of recall ratio are high. A very broadened search is to be conducted to satisfy his needs. This user's high recall requirements allow him to tolerate lower precision, as he doesn't want to miss anything.

On the other hand, a user who wishes to see some recent articles on a particular topic will expect high precision and less recall ratios in the output. In other words, he doesn't want to see everything on the subject but will expect the set of documents presented to him are highly relevant and he may even be particular about the currency of the items retrieved.

11.4.4 User Effort

It is to be noted that the more effort a user is willing to expend in exploiting an information storage and retrieval system, the better results could be obtained in terms of recall or precision ratios or possibly both. For example, in *delegated searches,* a user could put more effort at the search request stage through a detailed discussion of his requirements with a staff member of the system. In completing a detailed search request form he may include his requests in verbal form, i.e., the keywords that express the concept in the request, one or two known sources of documents relevant to the topic etc. He could examine the proposed strategy and suggest alternatives, if necessary, before the search is actually conducted.

The effort may also be put through the technique of iteractive searching where a partial or preliminary search is conducted and the user examines the output and indicates the relevant items. A revised search strategy may be prepared and executed to retrieve more documents of the type to be relevant and less of the type to be not relevant. This process called *Relevance feedback* is carried out by adding terms from documents found relevant by the user and deleting terms included in the documents that are found irrelevant.

In *non-delegated searches* where the end-user searches the system by himself, a good understanding of the system, its coverage, the syntax of the command language etc. on the user part are required to derive documents with higher recall and precision ratios. The user-friendliness of the interface plays a major role in this context.

11.4.5 Response Time

Time taken by the system to respond to a query put to it also plays a major role in evaluating information storage and retrieval systems. Users may consider that the receipt of a system response beyond a deadline, say five minutes, will be of no value. Yet response time is always secondary to the recall and precision requirements and the users will be willing to expend more time if they are sure that the system will satisfy their information requirements.

11.4.6 Form of Output

The form of output presented by the information storage and retrieval system helps the user to identify the relevant documents with ease. The output of a system may be brief bibliographic citations, citations with keywords, citations with keywords and abstracts or the complete text of the documents. The more information given to the user, the easier it is for him to make accurate relevance prediction. Similarly, the format of presentation is also important. If the descriptions are presented in a tagged format it may facilitate rapid scanning and helps the user to discard irrelevant items fairly easily, compared to continuous text format.

A study conducted by T. Saracevic indicated that out of 207 documents judged relevant from the full text, only 160 were so judged from abstracts and 131 from titles.

11.4.7 Assessing Recall and Precision Ratios

Establishing the performance figures of recall and precision ratios for a representative set of searches (test searches) will be helpful to identify examples of recall and precision failures i.e. situations in which known relevant items were not retrieved and situations in which known irrelevant items were retrieved. They arrived at figures which may be expressed in the form of a 2 x 2 table of search results as in Figure 2.

USER RELEVANCE JUDGEMENT

	Relevant	Not Relevant	Total
Retrieved	a "Hits"	b "Noise"	a+b
Not retrieved	c "Misses"	d "Correctly rejected	c+d
Total	a+c	b+d	Total collection a+b+c+d

Figure-2: 2 x 2 table of search results.

The total collection size (a+b+c+d), the total number of items retrieved (a+b), and the total number of items not retrieved (c+d) are the values directly observable. The two values, relevant and irrelevant items can be established by getting a feedback from the users where he judges some items as relevant (a) and some as irrelevant (b). The degree of relevance is judged by the user on some type of a scale as 'major, minor or no relevance. The user is asked to indicate reasons for his various judgments, which will be further, analyzed to improve system performance.

The two other values 'c' and 'd' (misses and correctly rejected) are yet to be established. The crude method to get these figures will be to present all the non-retrieved items (c+d) to the user for judging their relevance (c) and irrelevance (d). This method may be suitable on sample studies conducted on small prototype systems but is impractical in the case of large systems.

As it is impractical to assess the absolute recall values on large systems, the concept of best possible recall estimate is introduced. For example, the system is supposed to contain 'X' number of documents which can be judged as relevant if they are presented to the user. As this may be a large collection, a subset of 'X' i.e. 'X1' is selected by methods extraneous to the system under evaluation. 'X1' can be composed of relevant documents known to the user, or the documents found relevant by the evaluator through extraneous sources (e.g. other information centers or published indexes) or can also be comprised of partly of items from the first and second sources.

For example, a user has information on two relevant documents on hand when he makes a search request. A paraliel search by the evaluator on other sources may yield 12 possibly relevant items, as per the understanding of the evaluator of the topic. Of these 12 items, the user considers 8 relevant and it gives us 10 relevant documents (2+8) on the topic i.e. 'X1'. A search actually conducted on the system under evaluation retrieves 7 out of these 10 documents, then the recall estimate is (7/10) x 100 or 70%. The subset 'X1' is supposed to be a representative of the set 'X' and the recall estimate may also be attributed to collection 'X'. In this example, the total number of records relevant in the sample (a+c) is 10, the relevant records retrieved (a) are 7, and the number of relevant records missed (c) is 3.

Similar type of studies can also be conducted on a small sample of the database (source documents), say 1 year collection. A group of test searches may be conducted on this sample recall estimates can be established.

By above methods performance figures for a group of 'test searches' can be established and all the values in the 2 x 2 table of search results can be derived and further subjected to failure analysis.

11.4.8 Other Parameters Used in Evaluation

While most evaluation studies are based on the analysis of the two performance figures recall and precision ratios, the following parameters are also considered in some studies.

* *Fallout ratio* - the proportion of non-relevant documents retrieved in relation to the total number of non-relevant documents; also referred to as discard or the conditional probability of a false-drop

  Noise factor - the proportion of retrieved documents that are not relevant, the complement of precision ratio

* *Selectivity* - the proportion of non-relevant documents not retrieved i.e., correctly eliminated in relation to the total number of non-relevant documents

* *Specificity* - the proportion of relevant documents in relation to the total number of documents in the collection

* *Novelty* - the proportion of the retrieved documents that were not known to the user in relation to the total retrieved documents

11.5 FAILURE ANALYSIS

Failure analysis is diagnostic and is the most important part of the entire evaluation program. It involves a careful examination of the documents in the collection, the indexing policies and procedures, the indexing language(s), the original request posed to the system, the interpretation of the request by the searcher (in case of delegated searches), the search strategy used and the complete assessment of the output by the user.

Failure analysis conducted on a large number of searches yields large body of data. This data when analyzed and interpreted, clearly indicates the principal problem areas in the system - the areas that require attention and modification.

Failure analysis will attribute the recall and precision failures encountered to the principal subsystems and will identify the particular kind of failure occurring in each subsystem. The possible sources of failure may be identified at different stages of the system. They are:

1) User Expectations

User expectations on the coverage and appropriateness of the system (in terms of comprehensiveness and currency). Selection of an inappropriate system may lead to user dissatisfaction (e.g. Selection of BIOSIS instead of MEDLARS for topics on medicine)

2) Submitting the Search Request

Once the user identifies his information requirements, he has to submit it in terms of a verbal request. The quality of the request statement depends on:

User's interpretation of system capabilities and limitations:

* User's ability to describe his information needs in precise terms

* The degree of assistance given to the user by the system (e.g. carefully structured search request forms, interviews conducted with the user in assessing his information requirements, iterative search procedures followed, or some type of user training program).

3) Interpreting the Request

Once a request statement is submitted, it must be translated into a formal search strategy. The set of variables that affect the recall and precision ratios of a search at this stage include:

* The analyst/searcher's interpretation of user requirements (may be accurate or inaccurate)

* The ability of the vocabulary to describe concepts included in the request

The vocabulary may be having only higher level generic terms to express the concepts. e.g. A generic term 'Tissue culture' is only included in the vocabulary and the lower level terms such as 'cell culture, anther culture, shoot tip culture etc.' are omitted; A generic term 'arc welding' is only included, omitting the narrower term 'argon arc welding'.

* The ability of the searcher/analyst to recognize and cover all possible approaches to retrieval. E.g.: Omission of terms like cacao, coffee while a search is made on adverse affects of beverages may lead to less recall ratio

* The level and quality of search strategy adopted

As mentioned earlier, a broad search strategy may lead to high recall ratio with low precision, while a tight search strategy with stringent terms may lead to high precision but low recall ratio.

The quality of the search strategy also depends on the searcher's understanding of the syntax and semantics rules framed by the system in constructing the search statement, the precedence of Boolean/relational operators etc. Wrong usage of operators or illegal syntax may lead to low precision and recall ratios.

4). Searching the System

Once a search strategy is formulated it is to be matched against the contents of a database. Another set of factors which influence the system performance at this stage are:

* the *indexing policy,* particularly policy regarding exhaustivity of indexing expressed in number of terms used to describe the concepts covered in a document

* the *inaccuracy of indexing* i.e. omission of important terms or assignment of incorrect terms,

* the characteristics of the indexing language in terms of *specificity*

* the ability of the indexing language and its *syntax* in expressing the concept dealt within the document and the *context* in which the assigned terms are used

For example, sorghum is considered as staple food in African countries, while it is treated as a weed in wheat forms in USA. Consider the following title for indexing:

Breeding Wheat for resistance to Sorghum

The possible terms in expressing the concept could be:

Triticum vulgare (used for Wheat)

Plant breeding

Weed resistance

Sorghum bicolor (used for Sorghum)

In a pre-coordinated indexing system the subject string could be:

Triticum vulgare - plant breeding - weed resistance - Sorghum bicolor

which expresses the concept and context at which the above terms are used. Following the indexing rules at both stages: constructing an index string and a search strategy will result in high precision ratio in pre-coordinated indexing systems.

If the vocabulary and indexing methods used are not capable of expressing the concept and the context, false coordinations will lead to precision failures, particularly in post-coordinated indexing systems. For example, the document mentioned above may be retrieved against a search request -.*Breeding Sorghum for weed resistance* and a search strategy.

Sorghum bicolor AND plant breeding AND weed resistance,

as the above terms are included as keywords omitting the context in which they are used. In the above example, sorghum is considered as weed in wheat farms but this information is missing at indexing level.

5) Screening the Output

Before the search results are submitted to the user, they are usually screened to eliminate irrelevant items by the searcher or analyst. The screening operation is carried out to increase the precision ratio without effecting much the recall ratio. The success of screening process depends primarily on the accuracy of the analyst's interpretation of the user's requirements and secondarily on the quality of the document surrogate presented to the analyst.

The above discussion shows that a single factor: analyst's interpretation of user's requirements may affect the recall and precision ratios at different stages. The various sources of failure are cumulative and may result in a system that cannot operate very close to 100% recall or 100% precision.

In the end-user searches, one important source of failure - the analyst's interpretation of the user's needs - is eliminated. But the level of understanding of the system capabilities on user part and the user's inability to think of all possible approaches to retrieval may result in a large number of failures in end-user searches.

Failure analysis identifies the principal causes of failure in a particular system. Once the causes are identified, system modifications can be made that will reduce these failures in future.

11.6 IMPROVING THE SYSTEM PERFORMANCE

As mentioned earlier, an evaluation program is not conducted as an intellectual exercise, but is supposed to be diagnostic and therapeutic in the sense, the process should not only identify the causes for failure but should also suggest measures to improve the performance of the system. The cost of conducting an evaluation study can only be justified in terms of improved performance resulting from the investigation.

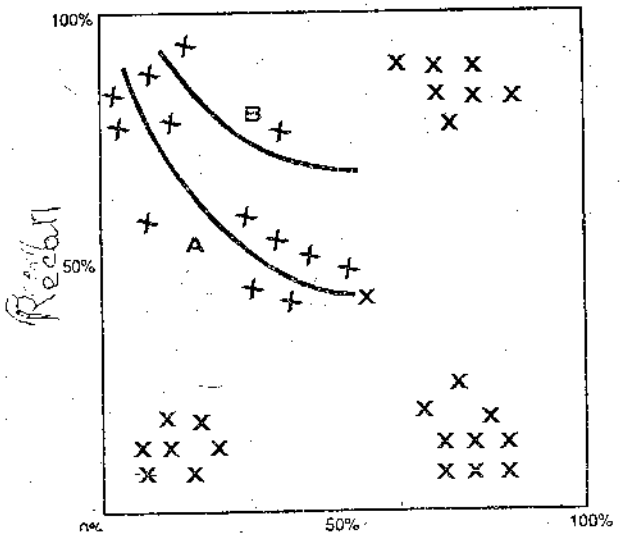


Figure-3 Scatter diagram of Search results.

In the above illustration, the individual performance figures (recall and precision ratios) for a group of test searches are marked by a symbol 'X'. The markings indicate that they scatter widely. There are some very good searches (marked 'X' on the top right-hand corner), some very bad results (bottom left-hand corner), some high recall-low precision results, some high-precision and low recall results and some compromised results in terms of recall and precision ratios. The individual performance figures can be averaged to arrive at an over-all average performance figure. By applying different search strategies to each of the test searches in the group, a series of average performance points can be derived and used to plot an average performance curve A for the existing system. The performance can be raised to curve B after necessary modifications made to the system at different subsystem levels. Recommendations may be made for modifications at different subsystem levels. Conducting another group of test searches on the modified system derives the performance figures for Curve B.

Some possible recommendations could be:

* making the user aware of the system scope, capabilities and limitations, its coverage through brochures, seminars, training programs etc.

* increasing the assistance provided to the user in formulating his search request

* Updating the vocabulary regularly with new terms that represent emanating new concepts in a subject area and establishing their relationships with the existing terms

* designing and implementing indexing procedures where the terms used to represent the concepts can also retain the context in which they are used and making necessary alterations to the software to meet these requirements

* grouping terms under major or primary descriptors and minor or secondary descriptors where primary descriptors represent the concepts dealt with in detail in the document and the secondary descriptors represent the concepts mentioned or not covered in detail in the document. Limiting the search to major descriptors leads to higher precision ratios while including the minor descriptors leads to higher recall. E.g.: The two titles below are to be indexed

Title *Introduction to Operating Systems"* deals with the concepts of operating systems in detail. A few chapters are included to introduce operating systems like "MS-DOS, UNIX" and the like.

The second title *Introduction to Unix* deals in detail with the specific operating system "UNIX".

The major descriptor(s) for the first title could be "Operating systems" and the minor descriptors "MS-DOS and UNIX", while "UNIX" becomes the major descriptor for the second title.

* switching to high speed hardware and software with better algorithms to reduce the response time

* improving the quality of query forms and help messages provided to the user by the user interface

Modifications made at the output end of the retrieval process (i.e. at the request stage - in interaction with the user, understanding his requirements and search procedures) can have immediate effect in improving the system performance than changes made at the input end i.e. indexing procedures and the indexing language.

Evaluation of an information storage and retrieval system could not be a one time study. It is to be noted that continuously monitoring system performance or conducting evaluation studies at regular intervals will help to identify specific failures and problem areas as they arise and to make necessary modifications wherever these are justified.

11.7 EVALUATION OF COST-EFFECTIVENESS

Cost-effectiveness refers to the relationship between level of performance (effectiveness) and the costs involved in achieving this level. It is designed to find the least expensive means to carry out a given set of operations or to obtain the maximum value from a given expenditure. There may be several alternative methods that could be used to obtain a particular performance level and these can be costed and compared. The costs that are relatively fixed (e.g. equipment purchase, developmental costs, costs involved in acquisition and indexing of the present database) and the costs that are relatively variable (based on the usage and mode of usage) are to be estimated. For example, if the number of searches conducted in a year increases by 50%, then the cost per search is considered reduced. Similarly, the mode of interaction with the user (personal visit, mail, telephone etc.), the mode of interaction with the system (on-line or off-

line), adding or eliminating screening operations, changing the professional level of the personnel conducting the searches, providing friendly user-interfaces and encouraging end-user searches thereby reducing the number of skilled professional dedicated to search services, all will contribute to variations in the cost per search.

A comparative study conducted on these alternatives may help to select the most promising alternative in terms of costs and effectiveness

The cost-effectiveness of an information system can be improved by either:

1) maintaining the present performance level while reducing the costs of the operating the system

2) holding operating costs constant while raising the average performance level

In an information storage and retrieval system, the basic alternatives and system tradeoffs relate to the input and document indexing operations on the one hand, and to the search and output transactions on the other. A particular performance criterion - for example, at given precision level - can normally be attained in many different ways, each of them involving different cost levels. Precision may be raised by using a highly specific indexing vocabulary but requires high indexer proficiency and large indexing costs, but reduces cost involved in relating to searching and screening efforts. Alternatively, the indexing may be performed more casually, but for high precision the output might be screened by experts, thereby decreasing costs at indexing level and increasing the same due to lengthened search time and screening time.

In some cases, it is possible to obtain quantifiable information, which relates to various systems' alternatives to the effectiveness or quality of the output product. The following relationships may be cited as examples:

1) collection coverage versus expected proportion of retrievals (inversely proportional)

2) indexing effort and time versus search effectiveness (directly proportional)

3) specificity of the indexing language and low recall and high precision ratios

4) Equipment complexity versus processing limitations (inversely proportional)

11.8 EVALUATION OF COST-BENEFITS

Cost-benefit evaluations are not easy to handle because the direct benefits of a retrieval service are difficult to identify and measure. Cost-benefit analysis requires a systematic comparison between the costs of individual operations and the benefits derivable from them. Yet, these benefits could be measured indirectly by

* comparing the cost of the service with the cost of obtaining the same by other means

* estimating the time gained or the increase in productivity resulting from the use of the system

* estimating losses owing to the lack of such a service

* estimating the benefits or avoidance of losses due to improve decision-making due to the availability of information from the system

* avoidance of duplication of research effort or projects that have either been done before or that have been proved infeasible by earlier investigators' simulation of invention

For example, a current awareness service that suggests possible new products, new applications for existing products, possible markets for industrial waste or less expensive methods of fabrications to industrialists might be justified economically.

11.9 LET US SUM UP

By the above discussion, it should be clear that most of the evaluation studies conducted on information storafe and retrieval systems have used recall and precision ratios. However, it is also obvious that a great deal of effort has been invested to develop other means of evaluation: fallout, selectivity, specificity, novelty, noise etc. It should also be clear that evaluation studies should always suggest modifications at micro-level i.e. subsystem level to improve the effectiveness of a system and these modifications are justified only when they are cost-effective and economically beneficial to the organization.

11.10 REFERENCES AND FURTHER READING

KING, D.W. and E.C. Bryant. *The evaluation of information services and products.* Washington: Information Resources Press, 1971.

LANCASTER, F.W. "Evaluatinn and testing of information retrieval systems" IN *Encyclopedia of. Library and Information Science.* New York: Marcel Dekker, Inc.; 1972. Vol.8; p.234-259.

LANCASTER, F.W. *Information retrieval systems - characteristics, testing and evaluation.* New York: John Wiley and Sons, 1979.

SALTON, Gerard and Michael J. McGill. *Introduction to modern information retrieval.* Auckland: McGraw-Hill International Book Co., 1983.

VICKERY, B.C. and Alan Vickery. *Information science in theory and practice.* London: Butterworths, 1987.

11.11 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) What is Evaluation ? Explain the different measures used in evaluating an ISAR system.

2) Discuss the advantages of Failure Analysis and suggest measures to improve the system performance.

II SHORT NOTES

a) Cost-effectiveness

b) Cost-benefit analysis

UNIT - 12 : EVALUATION OF *ISAR* SYSTEMS - EXPERIMENTS AND CASE STUDIES

Structure

12.0 AIMS AND OBJECTIVES

This unit aims to provide an overview of the various experiments and case studies carried out towards evaluating the ISAR systems.

After studying this unit, you should be in a position to

- discuss the experiments carried out on ISAR systems, popularly known as Cranefield Projects 1 and 2
- describe the evaluation of experimental systems and operational systems
- explain what is a user interface and the studies done towards evaluating the effectiveness of user interfaces.

12.1 INTRODUCTION

As the literature on evaluation of information storage and retrieval systems is voluminous and still growing, a sample of seven case studies are considered in this unit that can represent the efforts made in this area. The first case study CRANEFIELD I compares the efficiency of 4 indexing systems and the associated file organizations in terms of recall and precision ratios. It also considers the human factors that may lead to unsatisfactory results. In CRANEFIELD II, the components of indexing languages and the effects of these various components on overall system effectiveness are studied. Evaluation of FAIRS limited itself to the comparison of different search strategies on recall and precision values. The study conducted on MEDLARS considered different kinds of users and different levels of interaction between the user and the specialist, and the effect of these parameters on the overall performance of the system.

All the above systems use evaluation assigned indexes i.e. human experts analyze the content of the document and assign keywords with the aid of controlled vocabularies. The SMART case study considered the possibility of eliminating this human element in indexing and experimented with the possibility of providing systems with automated indexing and searching capabilities. A comparison is also made with fully manual MEDLARS system. The possibility of developing an intermediary search system or user interface based on the artificial intelligence and expert system methodologies, even though limited to specific subject areas, is explored in CANSEARCH study. The last case study, CONIT, considered the possibility of providing a common command language to the end-user so that he can access multiple databases in a networking environment and satisfy his information needs. A comparison of the interface CONIT and the human expert in arriving at satisfactory search outputs is also considered.

12.2 CRANEFIELD EXPERIMENTS ON *ISAR* SYSTEMS

12.2.1 Cranfield Project 1

The first extensive comparative test of information retrieval systems was undertaken at Cranfield, UK. The initial phase of the study, Cranfield I, was begun in 1957 and fully reported in 1962 by C.W. Cleverdon in his *"Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems"*. The project was financed by National Science Foundation.

The project attempted to test the effectiveness of four indexing technologies and associated file organization:

* an alphabetical subject catalog based on a subject headings list and a set of rules to guide the construction of headings

* a UDC classified catalog and an alphabetical chain index to the class headings constructed

* a catalog based on a special faceted classification, and an alphabetical index to the class headings

* a uniterm co-ordinate index controlled by an authority list of uniterms.

The study involved 18,000 documents and a total of 1200 manufactured or synthetic search questions devised after a study of each document. A search is considered successful when the strategy was able to retrieve the document in question.

The variables tested include:

* the systems as given above

* the type of document- half were research reports and half were journal articles chosen equally from a general field of aeronautics and from a more specialized field of aerodynamics

* qualifications of the indexer and his experience

* the indexing time in minutes per document

* the indexer's rate of learning the indexing policies and procedures

* the number of index terms assigned to each document

Interative searching methods (modifying the search strategy to get a desired output) were also used. The twin measures of performance, recall ratio and precision ratio were established.

The results of the study indicated surprisingly little difference in the performance of the four systems compared. Some of the findings include:

* all the four indexing systems were operating with an effectiveness of 60-90 per cent recall ratio i.e. uniterms with 82 per cent, alphabetical 81.5 per cent, UDC 76 per cent and Faceted classification with 74 per cent

* success in retrieving documents in general on aeronautics appeared to be 4-5 per cent greater than for the specialized field, aerodynamics, probably due to the lack of specificity in vocabulary in representing concepts in the specialized field

* no significant differences were observed in the case of retrieving documents indexed by different indexers

* increased time spent on indexing improved chance of recall

* the more experience the indexers gained in indexing, the more success in retrieval

* the source document was found in the first search attempt in 48 per cent of searches, in the second attempt in 35 per cent of searches, in the third attempt in 12 per cent of searches and so on.

Failure analysis conducted for a sample of 495 failures to retrieve the source document helped to identify 526 individual causes for failure at different subsystems. These were grouped as below:

1) Question failures (17 %)

   * too detailed (referring to only a small part of the document indexed)
   * too general (answerable by too many documents)
   * not easily understood
   * misleading
   * incorrect (the questioner had misunderstood the source document)

2) Indexing failures (60 %)

   * insufficient indexing (omission of important concepts, assigning too generic terms, misunderstanding the document)
   * omissions caused by shortage of time given to the indexer
   * overdetailed indexing (omission of general theme of the document)
   * careless indexing (wrong descriptors used)
   * incomplete entries (omission of rotated UDC entries, of cross references etc.)

Searching failures (17 %)

   * misunderstanding of the question by the searcher
   * failures to use all concepts in question

* insufficient searching
* incorrect searching (wrong entries checked)

4) System failure due to weaknesses in the descriptor schedules (6 %)

The study indicated that only 6 per cent of retrieval failures were attributed to the indexing system used; while human errors in indexing and searching lead to 60 per cent and 17 per cent failures, respectively, followed by user's ability in expressing his information requirements and the analyst ability in interpreting the user requirements to 17 per cent. From the study it was concluded that file organization is relatively unimportant in the performance of an information storage and retrieval system, while the specificity of the vocabulary and the exhaustivity of the indexing have a direct effect on the effectiveness of the system.

12.2.2 Cranfield Project 2

In the second phase of the Cranfield project, which begun in 1963 and fully reported in 1966, the components of indexing languages and the effects of these various components on the effectiveness of the system i.e. in terms of recall and precision ratios, were studied. The study was limited to a collection of 1400 documents (articles and reports) in a specialized field, high-speed aerodynamics and aircraft structures. Unlike Phase 1, (which was based on manufactured or synthetic questions), the second phase obtained genuine search questions/ requests previously submitted by the users.

Many devices are used while indexing a document in an information storage and retrieval system, some to increase precision and others to increase recall. A particular indexing terminology is a combination of such devices. For example, UDC uses the following devices to improve recall:

* grouping synonyms in its alphabetical index, e.g. air cushion vehicles- 629.137, ground effect machines - 629-137, and hovercraft- 629.137

* grouping word forms in its alphabetical index, e.g. weld - 621.791, welded- 621.791, and welding - 621.791

* linking broader or narrower terms in the schedule, e.g. 662- beverages, and 662.3 wines and

to improve precision:

* preco-ordination in the schedules, e.g. 533.6.071 - wind tunnels, 533.6.071.4 - wind tunnel instruments

* preco-ordination in the catalog, e.g. 669.71:621.791 - welding of aluminum

Cranefield project II studied the relative contribution of each of the devices on recall and precision ratios by varying each of these devices while holding others constant, and assessing the individual effects of a number of factors. As terminologies used in indexing

devices have a direct effect to performance figures, different indexing terminologies were constructed as follows:

1) A series of phrases were selected that express the concepts from the document. e.g. 'axial flow compressor', 'laminar boundary layer', etc. this become the first indexing language i.e. project II.1. The terminology is expanded to include synonyms and resulted in II.2.

2) All the concepts used in the test indexing (i.e. II-1) were organized in a hierarchical classification.

E.g.:	L43	GAS FLOW
	L44	INITIATED BY STRONG SHOCK
	L45	LOW DENSITY FLOW
	L46	RAREFIED GAS FLOW
	L47	FIRST COLLISION REGIME
	L48	MERGED LAYER REGIME
	L50	VISCOUS LAYER REGIME
	L51	NAVIER STOKES REGIME
	L52	DISSOCIATED STREAM
	L53	GAS EXPANSION

Expanded indexing languages based on the above hierarchical classification of terms were produced as follows:

II.12 species included, i.e. if the broader term was L46, its narrower terms L47-L51 were added

II.13 Super-ordinates included, i.e. to L46 and its narrower terms , the broader term L43 was added

II.14 Collateral included, i.e. to L46, its broader term, L43, narrower terms L47-L51, and collateral terms, L44, L45, L52, and L53 were added

Other indexing terminologies were produced by adding selected narrower terms or correlated terms, the selection being based on the context of the question asked.

3) The phrases selected from documents that represent the concepts in the documents i.e. II.1 were broken to create a list of uniterms. e.g. 'axial flow compressor' into 'axial', 'flow', and 'compressor' and these single terms formed indexing language I.1

4) The uniterms in indexing language I.1 were examined to determine: (a) synonyms, (b) words with common root, and (c) quasi-synonyms (terms whose meaning overlaps on occasion. E.g.: To the uniterm Flow,
flux, and stream are synonymous,
flowing is a word with a common root flow
and motion, movement are quasi-synonymous.

These three groups resulted in indexing languages I.2, I.3, I.5

5) The uniterms used in I.1 were organized into a hierarchical classification

E.g.: E74 COMPRESSORS
E75 JUMO
E76 MULTISTAGE
E77 FAN
E79 PUMP
E80 TURBINE

Expanded indexing languages were produced by three stages of hierarchical reduction, resulting in I.7, I.8, and I.9 indexing terminology as follows

I.7 to the narrower term, its collateral terms and the broader term were added - to E76, E75 (collateral) and E74 (broader) are added

I.8 to the terms in I.7, selected collateral terms of the broader term (i.e. compressors) were added i.e. to E76,E75,E74, E79 is added

I.9 to the terms in I.8, the remaining collateral terms of the broader term (compressor) were added i.e. terms E77 and E80 were added.

6) The conceptual phrases in II.1 were translated in to terms ived from EJC thesaurus. While this was indexing language III.1. It was expanded into: III.2 (including narrower terms only), III.3 (including broader terms only), III.4 (including related terms only), and III.6 (including all the three).

7) Uniterms selected from titles formed a separate index (IV.1), expanded to include word forms (IV.2) Uniterms selected from abstracts formed (Iv.3), expanded to include word forms (IV.4).

All together, 21 indexing languages were formed and compared to assess their effectiveness on recall and precision ratios.

Search strategies used may be illustrated as follows:

A. A question put to the system was broken into words, coordinated and matched against the words in the indexing language, I.1. In case, a specific document was not retrieved, coordination level was successively reduced by dropping one term at a time.

e.g. Breaking a question "determining the effect of chocking flow coefficient on compressor stage characteristics by test data analysis" results in nine terms: analysis, characteristic, choking, coefficient, compressor, data, flow, stage, test.

A relevant document in the collection that does not include the terms " analysis, data, test" in the index but dealing with the concept in the question could not be retrieved by co-ordinating all the 9 terms.

The method of dropping any one term at a time at each coordination level helped to retrieve the document at a co-ordination level of 6, where the question words "characteristic, choking, coefficient, compressor, flow, stage" would retrieve the document.

B. Words derived from the question were replaced by their synonyms (i.e. from I.2) and searches were conducted at different co-ordination levels.

Similar experiments were conducted for each indexing language, at each co-ordination level. A total of 221 questions were searched in each language and average recall and precision ratios were calculated for each language at each co-ordination level.

A "normalized recall" figure was calculated by summing up performance at all co-ordination levels.

The main conclusions of the study were:

1) By shifting from an indexing language using uniterms (I.1) to one using concept phrases (II.1), normalized recall fell from 65 to 45 per cent

2) As the single term language expanded (I.7-I.9), normalized recall fell from 65 to 61 per cent

3) Expansion of the concept language (II.2) into II.12, II.13, II.14 raised normalized recall from 45 to 57 per cent

4) Controlled terms from a thesaurus (III.1) showed a value of recall (62 per cent), which is less than that of uncontrolled single terms (I.1)

5) As thesauric relations were introduced into controlled terms (i.e. III.2, III.3, III.4, III.6), normalized recall fell from 62 to 59 per cent

6) Abstracts (IV.3 and IV.4) gave a higher normalized recall value than title (IV.1 and IV.2)

The study indicated that, taking both recall and precision into account, the index languages which used uncontrolled single terms (i.e. natural language systems, such as uniterms) performed better than other controlled vocabulary systems.

But as B.C. Vickery points out, in Cranfield Phase II, excessive attention was paid to a single factor, the indexing language, affecting retrieval performance. The project had not considered the factors such as:

* the contents of the system as affected by the input policy,

* indexing policy as affected by the estimate made of user needs

* indexing procedures as affected by a number of factors: the method used to select concepts, the experience of the analysts, the time given to them to index a document, the aids used in indexing etc.

* the specificity of the indexing language

* the devices the indexing language has for varying a search strategy

* the search purpose, the time allowed to the searcher, his search experience and available search aids

* file organization and search equipment.

12. 3 EVALUATION OF OPERATIONAL SYSTEMS

12.3.1 *FAIRS*

D.E. Berninger and his associates evaluated the operational system. FAIRS, the retrieval system of the U.S. Federal Aviation Agency. The system contains ten thousand technical reports indexed with the aid of a thesaurus of descriptors. Whenever a specific term was used, the corresponding generic term was added to the index i.e. the method of up-posting was applied. The system was put to evaluation by searching it to answer ten search requests, actually selected from genuine requests previously submitted by the users. User feedback on the relevance of the retrieved documents was collected and the precision ratios were established.

Then, two methods were used to calculate the recall ratio. In the first method, users were presented with a 10% random sample of the collection and were asked to identify the relevant documents in the sample against each test search request. This figure of relevant documents 'X1' was used to determine the possible number of relevant documents in the whole collection 'X' as 'X=10 x X1'

In the second method a supplementary test was conducted on 20 source documents. A set of completely synthetic searches were carried out on these 20 source documents. A search was considered successful if it recalled the source document and the recall ratio was derived on the basis of percentage of successful searches.

If a query includes generic term G1 and specific terms S1 and S2, corresponding thesaurus entries may be:

G1 with NT S3, S4, S5 etc.

G2 with NT S1, S6, S7 etc.

G3 with NT S2, S8, S9 etc.

The above example shows that the two specific terms S1 and S2 belong to other generic terms G2, and G3 respectively.

Four strategies were used to derive the performance 'figures: precision ratio, recall ratio (by.first method), and recall ratio (by second method):

* search strategy A included originally chosen search terms (whether specific or generic) i.e. G1, S1, S2

* search strategy B eliminated all specific terms and was limited to generic terms G1, G2 and G3

* Search strategy C included generic terms G1, G2 and G3 and eliminated all non-pertinent specific terms S3, S4, S5, S6, S7, S8, S9

* search strategy D co-ordinated generic terms with selected specific terms - G1 with any terms specific to G2 or G3, then G2 with any terms specific to G1 or G3, and so on.

The results of the study are presented in the following table:

Strategy	A	B	C	D
Precision ratio	59	35	38	45
Recall ratio (first method)	22	73	66	50
Recall ratio (second method)	70	90	80	75

The results indicated that efforts to raise recall resulted in the precision falls. The recall ratios derived by the two methods were not the same, yet they varied in the same way with precision.

12.3.2 MEDLARS

A larger system MEDLARS containing 70,000 biomedical articles was evaluated during 1966-67. On average 6.7 subject terms were used to represent the concepts in each article. The terms were selected from a thesaurus, MeSH (Medical Subject Headings), which consists of about 7,000 main subject headings that can be supplemented by sub-headings. Hierarchical searches are supported by the system.

A selection was made from the existing user groups that could (a) supply a certain volume of test questions; (b) cover all the kinds of requests made (categorized as on diseases, on drugs etc.); (c) include all kinds of users (academic, research, pharmaceutical, clinical, government etc.; and (d) vary according to the degree of user/system interaction (personal interaction, positive or negative or no local interaction).

The 21 user groups so selected provided 302 fully analyzable test searches. Search output, along with photocopies of the articles were provided to the requester for evaluation using the scale: H1-of major value, H2-of minor value, W1-of no value and W2-of unknown value. Precision ratios were calculated for over all precision (H1 + H2) and 'major value' precision 'H1" only.

Sampling techniques were used to establish overall recall ratio and 'major value' recall ratio for each of the 302 searches and these are then averaged to arrive at the following figures.

	OVERALL	MAJOR VALUE
RECALL RATIO	57.7%	65.2%
PRECISION RATIO	50.4%	25.7%

Each search was analyzed in detail, and failures in recall and precision were ascribed to the indexing language, to indexing, to user-system interaction, to searching, to computer processing. This analysis lead to a series of recommendations on upgrading system performance.

In about 23% of the 302 searches, a recall failure and in about 37 of the searches a precision failure were attributed to "inadequate user-system interaction" that resulted analyst/ searcher's inadequate interpretation of the information requirements of the user.

The four levels of interactions recognized are:

1) Personal interaction - the user visited a MEDLARS center and discussed his information needs personally with a system operator

2) Positive local interaction - a local librarian discussed the information needs before transmitting the request to MEDLARS center

3) Negative local interaction - a local librarian simply transmitted the request

4) No local interaction - the requester mailed his request directly to MEDLARS center

It was hypothesized that the first group of requests would give the highest performance but the results showed the interactive groups 1 and 2 performed worse than the neutral groups 3 and 4.

The success of the neutral groups is due to the fact that the requester submitted his information needs in verbal form, in his own natural language, without being influenced by the logical and linguistic constraints of the MEDLARS system, as evidenced by no interaction with the system.

The failure of the interactive groups is due to the fact that the user has initially a less well-formed idea of what he is seeking (i.e. of the scope and constraints of the search) and when this somewhat imprecise need is discussed with a search analyst, in terms MeSH, it tends to become forced into the language and logic of the system. The final 'request' rather than representing what the user wants, represents what he thinks of the system can give him. It appears that little knowledge of the system on user part will lead to failures. Either the user should give full freedom to the searcher to analyze the search request and formulate a strategy or he should himself learn the technique of searching a system.

Lancaster commented that ".It appears crucial to the success of a MEDLARS search that the requester be required to write down, in his own natural language, exactly what type of literature, he is looking for."

12.4 EVALUATION OF AN EXPERIMENTAL SYSTEM - *SMART*

The SMART (System for Mechanical Analysis and Retrieval of Text) project was initiated in 1961 with an emphasis on fully automated procedures for the analysis, search, and retrieval of natural language texts and has become operational in 1964. From its inception, the system was designed both as a retrieval tool and as a vehicle for evaluating the effectiveness of a large variety of automatic search analysis techniques.

12.4.1 Design of SMART System

SMART system consisted of three parts: an automatic content description (indexing) system; a supervisory or monitory system; and an evaluation system. The SMART indexing system was based on seven language analysis tools:

* methods for automatically extracting important words from natural language texts of incoming user queries and documents excerpts (titles, abstracts or full texts);
* sophisticated suffix cut-off procedures which would be used to transform the words extracted to word stem form;
* synonym dictionaries or thesaurus;
* hierarchical term arrangement systems;
* syntactic analysis systems;
* semantic analysis systems;
* and statistical frequency analysis systems.

In SMART, a document or query is represented by a vector or terms i.e. words that carry the concept of document or query.

Steps in Automatic Indexing of Document or Query

1) The document text or query is broken into words

2) High-frequency function words such as "and, of, or , but, when, where etc." are removed with the help of a stop word list

3) The scope of the remaining word occurrences is broadened by reducing each word to a word stem using suffix removal procedures

E.g.: economist, economists, economical, economically, economize, economizes, economized, economizing, economies, etc. into ECONOM

4) multiple occurrences of a given word stem are combined into a single term for incorporation into a document or query vector (frequency analysis is carried out)

E.g.: document 1 dealing with fruits is represented in the following vector:

(apple, 4; pear, 3; guava, 2; plum,1)

which means that the document deals more with apple than the other three fruits

5) Transforming the word stem vectors into useful term vectors by two manipulations: first a term weight can be assigned to each term reflecting the usefulness of the term in the collection environment; and second, terms whose usefulness is inadequate as reflected by the low term weights can be transformed into better terms.

Terms weights are assigned in the following manner:

a) calculating the frequency of the term in a document or query.

b) calculating the discrimination value of the term in distinguishing the specific document from other documents in the collection

c) calculating the document frequency i.e. the number of documents to which the term is assigned.

$$\text{WEIGHT} = \frac{\text{TERM FREQUENCY}}{\text{DOCUMENT FREQUENCY}}$$

OR TERM FREQUENCY x DISCRIMINATION VALUE

6). Terms whose weights are neither too large or too small are incorporated directly into the document or query vectors.

Terms whose weights exceed a given threshold level are considered too broad and unspecific. These are rendered more specific by being combined with other terms into term phrases that contain two word stems before including into the vectors.

E.g.: people in need of information require effective retrieval services (original sentence)

PEOPLE INFORM EFFECT RETRIEV SERVICE word stems

PEOPLE INFORM	EFFECT RETRIEV	INFORM EFFECT	term
EFFECT SERVICE	INFORM RETRIEV	RETRIEV SERVICE	phrases
INFORM SERVICE			

In forming term phrases , the indexing system follows certain rules such as the word distance should not exceed 4 and these should be in a single sentence etc.

Terms with low weight are considered too specific and are broadened by grouping them into term classes similar to a thesaurus entry. The thesaurus class identifiers are then incorporated into document and query vectors, instead of the individual rare terms.

Documents in SMART system are automatically classified based on vectors and placed in clusters where items that appear reasonably similar to each other are placed in close proximity.

The supervisory or monitoring system could process the query and document vectors calling necessary language analysis tools and by suitable matching operations, could supply to the user, references to those documents whose content vectors appeared to be similar to the corresponding query vectors. The evaluation system provides formal assessments of system effectiveness in terms of satisfaction of users.

12.4.2 Search Process in SMART

Information is retrieved by a complete vector matching method providing for each query-document pair a coefficient of similarity. A ranking is obtained for the stored items (i.e. 100% similar, 90% similar etc.) in decreasing order of similarity, and available number of documents are presented to the user. This permits the user to consider first those documents which appeared to the system to be most similar to the specified query and select some relevant documents.

A new search operation could then be initiated by automatically altering the initial query vector so as to retrieve more documents similar to the documents considered relevant by the user. A number of such relevance feedback operations could be carried out so that user's information needs could be met satisfactorily.

12.4.3 Evaluation of SMART

A number of tests conducted in varied subject areas like engineering, aerodynamics, and documentation indicated that:

* the order of merit is generally the same for all three subject areas
* the use of unweighted terms is less effective than the use of weighted terms
* the use of document titles alone is always less effective for content analysis purposes than the use of abstract

* the thesaurus processes involving synonym recognition perform more effectively than the word stem extraction method, where synonyms and other word relations are not recognized

* the thesaurus and statistical phrase methods are substantially equivalent in overall system performance

* other dictionaries including term hierarchies and syntactic phrases exhibited poor performance

12.4.4 Comparison of SMART with MEDLARS

One of the aims of the SMART project had been the comparison of fully automated text processing systems with the manual indexing systems like MEDLARS. MEDLARS system is based on a manual analysis of documents and incoming search requests.

A sub-collection from the full MEDLARS collection and a subset of original queries submitted to MEDLARS were used for a comparative study, by processing both as per SMART methodology. The recall and precision results, averaged over 29 queries, exhibited recall ratios about 40% lower for SMART, than for MEDLARS; the precision loss was between 30 to 40%, where SMART used its standard word stem extraction method only. When the ranked outputs provided by SMART were used, the situation improved drastically, i.e. a deficiency of only 16% in recall and 19% in precision. Through the use of relevance feedback methods, this 15-20% deficiency in recall and precision turned into an advantage of 4-7% after one feedback operation, and of 10-13% for two feedback interations, over MEDLARS.

To improve the effectiveness of the SMART system an automatically generated dictionary that contains a list of all terms in decreasing order of term discrimination value, designed to exclude all high frequency terms was used. The automated dictionary provided a 10% improvement in recall and a 20% in precision, over the standard word stem extraction method. With the use of SMART thesaurus, improvements of about 25% in average recall and precision ratios were achieved over the standard word stem process.

The SMART-MEDLARS comparison can be concluded as follows:

1) The strong points of the automatic retrieval system appear to be the vector matching techniques which furnish ranked document output, the automatic construction methods for word control lists, and the feedback operations.

2) The simple word stem extraction process using document abstracts and query texts is only 15-20% less effective than the best available manual indexing based on controlled vocabularies.

3) Automatic language normalization procedures can be used to build dictionaries and thesauri, whose operations produce output results equivalent to standard manual indexing.

4) The SMART relevance feedback procedures produce large improvements in retrieval effectiveness.

5) The Boolean search techniques are inferior to vector-matching techniques that produced ranked output in decreasing query-document similarity order.

12.5 EVALUATION OF AN EXPERT SYSTEM BASED USER INTERFACE TO *MEDLINE - CANSEARCH*

A novel intermediary system (user interface), CANSEARCH, has been designed to provide access to cancer therapy literature on the MEDLINE database for doctors with no knowledge, training, or experience in information retrieval. The system is limited to the generation of legitimate search statements to query the MEDLINE database on a single specific subject area, cancer therapy. This constraint follows the expert system maxim of choosing a well-bounded domain where specific knowledge with respect to the domain can be applied to raise the level of performance.

12.5.1 Design of the System

The design philosophy is based on: the choice of a particular subject area, so that subject knowledge may be incorporated into the system; an abstraction of search space, to present the potential of the system in a way which enables efficient term searching; and the use of hierarchical menu selection techniques to minimize or eliminate the need of typing.

The controlled vocabulary of database is considered an abstraction of the contents of all the documents in the database. The knowledge base concerns general knowledge of clinical cancer therapy, knowledge of the controlled vocabulary of terms used for indexing cancer therapy documents in MEDLINE, and knowledge of specific indexing instructions.

Abstraction of the search space is achieved through two ways:

a) establishing a link between an indexing term from the domain of interest, one of the terms from Medical Subject Headings (MeSH) used in MEDLINE, and the documents to which it is assigned. e.g. While the term 'breast' is used as an indexing term in the domain, the equivalent term in MeSH 'breast neoplasms' representing breast cancer is selected through links established, while formulating the search request to MEDLINE

b) arranging the index terms into hierarchy such that the top level presents the overall domain of the system to the user.

Menu selection is achieved by a user pointing at the screen by a finger touch and the selected option is displayed in reverse video confirming selection. A selection in a hierarchical menu (situation) may lead to a sub-menu or with a set of controlled vocabulary of terms (action). Actually, it is the selection of terms from controlled vocabulary that results in the search strategy formulation.

IF (situation) THEN (action) rules embody the hierarchy of menus, the MeSH vocabulary, the rules of indexing, and search statement formulation.

A sample test query "The use of fluorouracil in the treatment of breast cancer", may take a search path as below by selecting options from the menus as presented to the user

* cancer at a particular site in the body AND therapy (from top level menu)
* specific primary sites (from sub-menu)
* breast from the controlled vocabulary displayed in a hierarchical form
* chemotherapy from top level therapy menu
* particular antineoplastic drugs/drug classes from chemotherapy sub-menu antineoplastic antimetabolites from drug classes sub-menu
* fluorouracil from antineoplastic antimetabolites term list (drugs list)

The final legitimate query statement generated by CANSEARCH in terms of MeSH would be

BREAST NEOPLASMS AND FLUOROURACIL

CANSEARCH assumes that the searcher, in this case a doctor, has a thorough knowledge of his subject area, else passing through various levels of hierarchical menus and selecting appropriate sub-menus and terms therein could not be handled by them with ease.

A sample of test searches conducted indicated that:

* novice end-users could use CANSEARCH to specify the subject of a query concerning cancer therapy
* CANSEARCH would be able to generate legal MEDLINE search statements
* relevant documents could be retrieved from MEDLINE using the search statements generated by CANSEARCH
* CANSEARCH would be as effective as a human intermediary or search analyst

12.6 EVALUATION OF THE EFFECTIVENESS OF A USER INTERFACE-*CONIT* COMMON COMMAND LANGUAGE

A user's information needs, on some occasions, may not be met by limiting his searches to a single information retrieval system. As and when his information needs vary, he has to select an appropriate retrieval system that can satisfy his information needs, particularly in a networked environment where a user is provided access to multiple systems.

Several heterogeneous retrieval systems are commercially available and there are significant differences among these systems in terms of command languages, search aids, (thesaurus etc.) provided by them. The end-user has to learn the syntax and semantics of each of the information retrieval system if he wishes to exploit its capabilities. Expert systems based interfaces like CANSEARCH are designed and implemented but they are confined to specific subject areas and available on specific systems only.

CONIT is a general purpose interface aimed at connecting three different commercially available retrieval systems (MEDLINE, SDC ORBIT AND DIALOG), which together contain 300 databases by 1983. CONIT connects the user to these three systems but presents to the user what appears to be a single, common (virtual) system by allowing user requests in a common command language. These requests are in turn translated by the interface into appropriate commands acceptable to the host system selected by the user. The interface provides instructions and additional search aids to help the novice user.

12.7.1 Automated Keyword / Stem Searching

The problem of effective searching by a novice user across databases with heterogeneous indexes was met by a natural language, free vocabulary approach to searching that emphasizes the use of keyword stems as the basis for searching.

A search on the topic 'transplantation rejection' results in two word stems 'transplant and reject'. Then CONIT conducts truncated searches on each of the stemmed forms in all the indexes that can be searched with a single command in the connected database. The sets retrieved for each individual subsearch are then combined with a Boolean OR and finally these separate unions are combined with the Boolean AND operator to yield the resultant set.

Searching in an information retrieval system is not limited to subject searching only. Users may wish to search the system on personal names, but the rendering of personal names varies from system to system. e.g. Lancaster, F.W. in one system and Lancaster, F W in another system.

To get over this problem, the user is permitted to request personal names searches in a common format. CONIT then translates this format into the one appropriate for the database being searched - correct spacing and punctuation between entry element and initials are supplied by CONIT.

CONIT names the subsearches and the resultant search and reports to the user the number of documents in each set. All this is done automatically without user intervention. If any of the subsearches yield null results, CONIT suggests browsing the index terms or the thesaurus around the non-responsive term. If a truncated search causes a search buffer overflow, CONIT replaces truncated search with an exact match, full-word search or full-phrase search.

12.7.2 Search History and Reconstruction

CONIT system has a search history recording and reconstruction capacity. For each search CONIT records the full search formulation, the database system used, the number of

documents found in the resultant set or in any component sets formed in creating the resultant set, and the set names as given by CONIT and by the retrieval system, and whether the set is currently available in the retrieval system etc. all this information will be available on-line to the end-user.

When a user requests any component or compound search formulation to be repeated in any other database or set of databases, CONIT refers to the search history and repeats the search, after connecting to appropriate systems and databases. The dropped or no longer available sets, in the original request are generated first before any operation (output generation or combining sets) is performed, totally transparent to the end-user.

12.7.3 Evaluation of CONIT

Some 16 end-users selected from different levels (two medical doctors, one non-academic university staff, two professors, one post-doctoral fellow, 6 graduate students, and 4 under graduates), none of whom previously operated either CONIT or any one of the connected retrieval systems, performed searched on 20 different topics using CONIT with no assistance other than that provided by the interface. These same users performed searches on the same topics with the help of a human expert who searched the retrieval systems directly.

The parameters considered for the experiments include: total search time, the time spent by the users in getting help from CONIT, the actual search time which includes the time spent in issuing commands and getting their responses, the time spent in displaying retrieved records and assessing their relevance, the number of relevant documents retrieved, the estimated number of relevant documents in the documents, and the number of databases searched.

The results indicated that:

* sometimes CONIT and sometimes the human expert were clearly superior in terms of search effectiveness i.e. recall and precision ratios. In general, however, end-users searching alone with CONIT achieved higher on-line recall at the expense of longer search sessions.

* in terms of search time which has a direct bearing on the cost of search, particularly in the case of commercial databases, experts had spent at least 20% less time than their counterparts, end-users.

* the number of relevant records found and viewed on-line was much higher for the user CONIT sessions than for the human expert sessions; this lead to longer search sessions in case of user CONIT sessions

* human experts seemed to be more sophisticated, complex and comprehensive in their search i.e. they used all possible tools to raise recall and precision

* human experts regularly took advantage of such precision-enhancing devices as proximity searching, important term searching, and subheadings and other controlled vocabulary searching

* experts also used such recall-enhancing devices as truncated searching and searching on all more specific terms given for a term etc.

* end-users did not make use of these devices, as they are not aware of their availability in the system

* a number of end-users used the facility of browsing indexes and thesaurus using terms found in document records, as CONIT suggests this during execution

It is concluded that advanced experimental intermediary techniques are capable of providing search assistance whose effectiveness is at least similar to that of human experts in some contexts.

12.7 LET US SUM UP

Let us recapitulate what has been discussed so far in this unit.

* A number of experiments and case studies have been carried out to evaluate the effectiveness of information storage and retrieval systems.

* The first study, popularly known as Cranefield Project 1, compared the efficiency of four indexing systems, while in the second phase (Cranefield Project 2) a number of variables studied.

* Evaluation of operational systems (FAIRS and MEDLARS), and experimental systems (SMART) have been discussed.

* A user interface has been devised, as an intermediary system, to help the end-users without any experience on searching by various information retrieval systems. The evaluation of interfaces, CANSEARCH and CONIT have been discussed.

12.8 REFERENCES AND FURTHER READING

1) VICKERY, B.C. 1970 *Techniques of information retrieval.* London: Butterworths, 1970

2) *KEY papers in the design and evaluation of information systems/* edited by Donald King. New York: Knowledge Industry Publications, Inc., 1978.

3) POLLIT, S. 1987. *"CANSEARCH: an expert systems approach todocument retrieval". Information Processing & Management* 23(2): 1987. pp.119-138

4) CHARAMELLA, Y. and B. Defude. A prototype of an intelligent system for information retrieval: IOTA. *Information Processing & Management* 23(4): 1987. p.285-303

5) BORGMAN, C.L., Case, D.O. and Meadow, C.T. "The design and evaluation of a

front-end user interface for energy researchers". *Journal of the American Society for Information Science* 40(2): 1989. p.99-109

6) MEADOW, C.T., Wang, J. and Yuan, W. "A study of user performance and attitudes with information retrieval interfaces". *Journal of the American Society for Information Science* 46(7):1995. p.490-505

7) HANCOCK-BEAULIEU, M., Fieldhouse, M. and Do, T. "An evaluation of interactive query expansion in an online library catalogue with a graphical user interface" *Journal of Documentation* 51(3): 1995. p.225-243

8) MARCUS, R.S. "An experimental comparison of the effectiveness of computers and humans as search intermediaries". *Journal of the American Society for Information Science* 34(6): 1983. p.381-404

9) SALTON: G. and McGill, M.J. *Introduction to modern information retrieval.* Auckland: McGraw-Hill International Book Co., 1983.

10) VICKERY, B.C. and Vickery, A. *Information science in theory and practice.* London: Butterworths, 1987.

11) ELLIS, D. *New horizons in information retrieval.* London: Library Association, 1990.

12.9 ASSIGNMENT

Critically examine the effectiveness of the indexing systems used in your library, documentation or information centre.

12.10 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) Write an essay on Cranefield Experiments on evaluation of various indexing systems.

2) Critically examine the evaluation studies on experimental and operational systems: SMART, FAIRS and MEDLARS.

3) What is a User Interface ? Discuss the efforts made towards evaluation of the effectiveness of User Interfaces.

II SHORT NOTES

a) COINT

b) SMART Vs. MEDLARS

BLOCK - IV : INFORMATION ACCESS AND RETRIEVAL

Computers and communication technologies have come in handy to tackle the problem of unprecedented growth of information. The advances in information technologies - storage media, software, graphics, networking, etc. have made considerable impact on access, organisation and delivery of information.

Though databases have been developed initially as byproducts of the various major print-based abstracting and indexing services, the publishers of such products have made them available to the users as online databases and CD-ROM databases. There are over ten thousand databases that are publicly available now to serve the academic, scientific, business and management communities. The scholarly literature from the researchers has been made instantly accessible through full-text electronic journals from the online databases. As no single service can guarantee to offer full coverage of the required literature, multiple database searching has become necessary.

Now the Internet has become one-stop-shop for all information services - e-mail, listservs, online discussion groups, searching the OPACs of university libraries, fulltext and bibliographic databases, etc. Information from government, educational institutions, research organisations, business firms, industries, banks, non-government organisations and individuals is now available with a click of a mouse button.

In recent times, library and information centres have started showing a lot of interest to improve their services through Expert Systems. Library Expert Systems have been designed to assist the librarians and information scientists in making certain decisions in problem-solving situations faced by them not only in library operations, but also in automatic indexing and abstracting, online intermediaries and expert system search engines.

In the present block, there are four units, viz.,

Unit-13 : Information Access -. Online and CD-ROM databases

Unit-14 : Database Searching; Search Strategies

Unit-15 : Searching the Internet

Unit-16 : Library Expert Systems.

UNIT - 13 : INFORMATION ACCESS : ONLINE AND CD-ROM DATABASES

Structure

13.0 AIMS AND OBJECTIVES

Online and CD-ROM databases now commonly operate side by side in libraries and information centres as media for computerised information delivery. In this unit we introduce you to the information access through major online and CD-ROM databases.

After studying this unit, you should be able to

* define term 'database' and explain the concept and types of databases
* explain online information services and list out various online service vendors and databases

* discuss the advantages and disadvantages of CD-ROM format and list out various check points used in choosing a CD-ROM publisher
* describe the features of major online and CD-ROM databases.

13.1 INTRODUCTION

Institutions worldwide spend billions of dollars annually on research and development. In this process they generate vast amount of Information that need to be accessed, managed and used effectively. Academic and scientific research funded by the governments adds substantially to this volume. Traditional methods of information management in the form of printed journals, books, reports, etc., contribute to delays in effective communication and causes loss of time and money. Increasingly, scientists are willing to abandon the traditions and pursue new technologies, in an effort to manage the information more effectively.

If the information need demands 3Cs, namely, Current, Comprehensive and Cost-effective, then one should have a mix of technologies, namely, On-line and CD-ROM.

The developments in the information technology, which include, computers and communications, have lead to this trend. The rapid advancements in computers, storage media, software packages, graphics technology for processing information, information networks, etc., have eased the situation considerably. These developments made it possible to convert large volumes of information into machine-readable form as databases and are accessible through telecommunication lines remotely, in the form of CD-ROMs and On-line accessing of information.

13.2 ELECTRONIC DATABASES

Computers and communication technologies come in handy to tackle the problems of unprecedented growth of information. Databases were initially developed as by-products of print versions of major abstracting and indexing services by the publishers and made them available to the users. Now they became prime sources for getting information instantly even from the remote places.

13.2.1 Database - Concept, Definitions and Types

The term, 'database' has several definitions. According to Computer Science community, a database is an accumulation of interrelated data or information, well organised into machine-readable records for easy retrieval by a data processing system. In other words, databases are computer-based systems, which collect data from a variety of sources and integrate them into consistent sets, thereby, to facilitate access, manipulations and corrections of the data for a group of potential users. The database concept is made operational by a Database Management System (DBMS), a software system which reforms the functions of creating and updating files, retrieving data and generating reports.

In an information retrieval environment, a database is a collection of information that is organised and searchable (and perhaps also updateable) in the context of one or more specific applications. For example, at the application level, such as finding books by a given author

from a catalogue database. The electronic databases hold key to improving information accessibility — the ease and quick with which our academic, scientific and management community can access information. Beginning with the Library of Medicine with its MEDLARS database system in 1960, the number and size of databases has grown exponentially, from 25 in 1965 to 9550 in 1995. Current estimates indicate that there are more than 6000 publicly available electronic databases accessible through an online vendor or batch processor as well as those available on CD-ROM, magnetic tape and diskette.

There are a variety of electronic databases. For our purpose, we can group the electronic databases into the following five types :

Reference databases
Full-text databases
Directory-type databases
Factual and Numeric databases
Pictorial databases

Bibliographic and numeric databases have revolutionised the way the information is handled. Given the capability of computer and communication technologies to store, manipulate and transmit vast amounts of information at very high speeds with greater accuracy, it has been possible to reduce the time factor involved in the information dissemination process. This, in turn, made possible to access more current information to the users than what comes through printed media. Although, they share the same purpose with their printed versions, electronic databases offer more complete, more intensive, more current and faster responses. The flow of information from print to database production and use is shown in Figure-1.

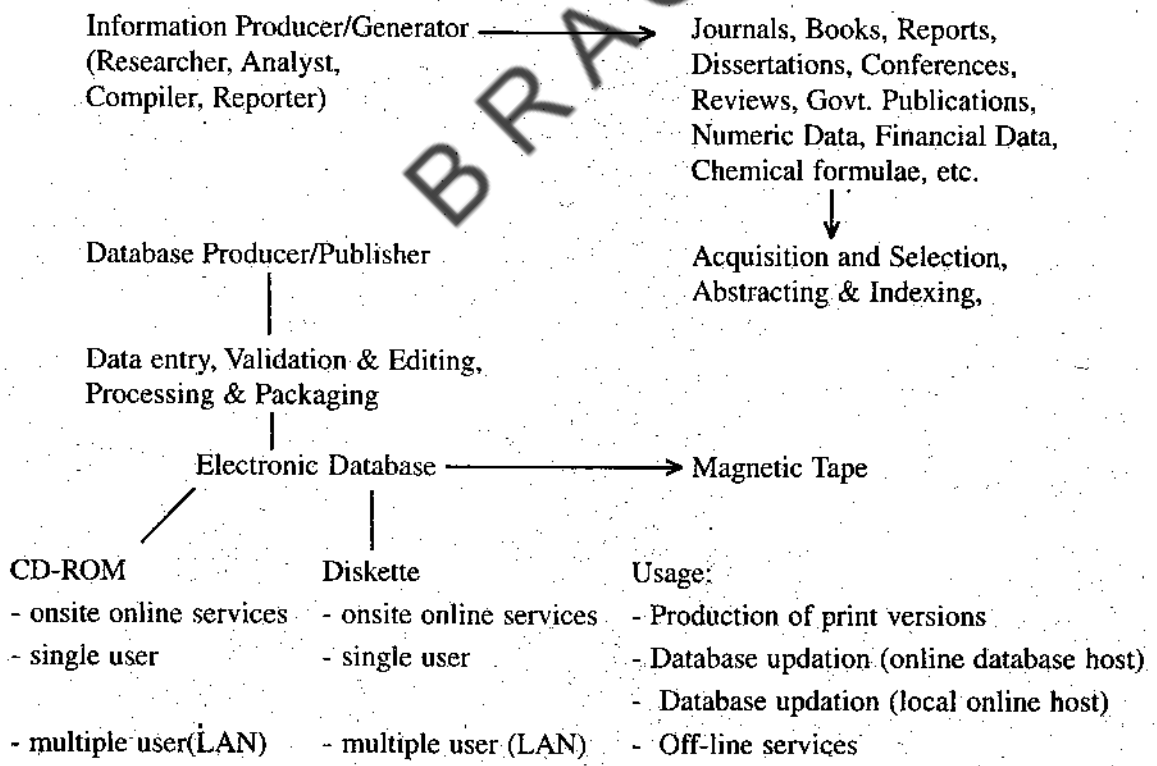


Figure-1: Electronic Information Path

13.2.2 On-line Services

A number of databases are available through online information service vendors. Through online information retrieval services, a user can get his required information from the databases in an interactive, conversation mode sitting at a remote terminal, even sometimes thousands of miles away from the host computer. The computer terminal is connected with the host computer by a telephone and telecommunications network.

Generations of On-Line Services

Broadly, three generations of online services are observed. They are

First Generation : Host-based
Character-based. Eg: Genie, CompuServe, BIX, Internet (DIALOG, The WELL)

Second Generation : Grapphics-based (Eg: Prodigy, America Online, eWORLD)

Third Generation : Agent-based
Client/Server based (Eg: AT&T Personal Link, IBM Intelligent Communications)

The evolution of on-line services has mirrored general computer technology from host-based to client/server, and from character-based to graphics and now to agents.

In the first generation services, information came from the host, and all the work was done there except for local screen-rendering and locally maintained address books, message folders, and down-loaded files.

The second generation services introduced Graphics User Interfaces (GUIs) and the exchange of graphics primitive; more computer power was presumed to be on the users' side of the pipe, and the architecture shifted subtly toward client/server.

In agent-based services, roving software objects can conduct business on behalf of users even when they are not connected to the network.

Major Online Services

There are several commercial online information service vendors and they offer their online services. Some of the major ones are -

1) Bibliographic Retrieval Services (BRS) Incorporation

2) DIALOG Information Services Inc.

3) National Library of Medicine (NLM)

4) Systems Development Corporation (SDC) Search Service

5) ESA - IRS

6) BLAISE

7) INFOLINE

8) Data-Star

9) OCLC

10) STN

11) RLIN

12) ORBIT

13) ECHO (European Community database)

13.2.3 CD-ROM Databases

CD-ROM stands for Compact Disc - Read Only Memory. It is 12 cm in diameter and has a thickness of 1.2 mm. It is a method of using currently existing optical technology to store a large amount of data, on a compact disc

The basis of technology is exactly the same as we find with audio CDs. A CD-ROM disc is composed of a very thin sheet of a metal, with a spiral track on it, going from the centre out to the edge, like the Vinyl LPs. During mastering and replication of the discs, a laser will burn a hole into the track or it will not, resulting a series of 'pits' and `land', which in turn are interpreted by computer system as 0s and 1s (binary system). Therefore, the data is written to disc digitally and read optically. A robust plastic coating, to ensure that the fragile metal surface does not get damaged, protects the metal surface. One side of the CD-ROM will have the label printed on it, while the other blank side contains data itself.

The exciting features of CD-ROM are its ability to play a wide repertoire of material, viz., books, periodicals, directories, educational materials, games, movies, music, etc. The communication modes/data types in a CD-ROM include : animation, graphics, software, sound, text, and video. CD-ROMs hold up to 660 megabytes of information (i.e. over 470 floppy discs of 1.44 mb), up to 18 hours sound, up to 700 million characters of text, up to 74 minutes of movies or other video.

CD-ROMs are cheap to produce and use. It costs only a few dollars per disc to duplicate a CD-ROM title in a smaller number. In high quantities (11,000 or more), the cost goes down to less than $1 per disc. With storage capacity equivalent to 250,000 pages of paper, publishing on CD instead on paper can save up to SIX TREES for each disc published.

Advantages of CD-ROMS

The major advantages of CD-ROM are :-

1) Permanent storage of data, that cannot be erased, scratched or mutilated like a printed book
2) Durability : shelf life upto 100 years
3) Portability : light weight and easy to mail in an envelope
4) Low cost : production cost is volume-based like printed books
5) Unlimited use : a single disc can be shared by multiple number of users
6) Data protection : the surface containing data is protected by a plastic coating
7) Ideal for libraries : shelf space is saved and low maintenance cost
8) Networking and data exchange: CD networks
9) Multimedia: able to combine text, graphics, sound, video and animation

Disadvantages of CD-ROMs

The major disadvantages of CD-ROMs are :-

1) Initial price in the form of subscription to the database producers is very high
2) Currency of CD product is delayed
3) Access is limited to one person per CD installation
4) Hardware and software compatibility and interchangeability still a problem
5) Searching is slow
6) Lack of standard retrieval software compels searchers to learn different systems
7) Multiple file searching, cross tabing is still a difficult job
8) Only a limited number of Indian publishers are bringing their publications in CD-ROM form

CDs with their enormous storage capacity, facilitate a variety of Multimedia and interactive applications. The applications are available on a variety of popular operating systems, such as Windows, Unix, OS/2 and Mac. The cross-platform developments allow the applications developed on one platform can be used on another platform. They support popular standards such as OLE (Object Linking and Encoding) and ODBC (Open Database Connectivity).

Another major development CD-ROM technology is that of CD Networks. Networks offer access to different CD-ROMs in the same campus. The major advantages include quicker and easier updating of software applications, enhanced security, and greater flexibility. The CD networking is made possible through a networkable automatic loading mechanism or Jukebox. It provides large disc storage capacity and multiple drives (4 to 6 drives are linked together internally). Software like Opti-NET, CD-Net, CD-PlusNet, MultiPlatter and SCSI Express provide

networked access to CD-ROM databases and permit sharing of information. The popular CD-ROM services like *Business Periodicals* Ondisc (BPO), a full-text database, allow networking to retrieve and print articles from the remote workstations.

CD-Recordable (CD-R) is a recent development, which has broken the monopoly of CD service providers using the elaborate techniques to 'master' and 'replicate' the CDs. The CD-R technology allows even the desktop PC users to publish their own databases and make limited distribution of their CDs in-house. Initially, it promoted many institutional libraries to produce CDs with OPAC data and distribute it to their users.

13.2.4 ON-LINE *vs* CD-ROM Databases

Let us compare the on-line systems with that of CD-ROM

	ON-LINE Systems		CD-ROM Systems
1)	It is basically an Access technology	1)	It is basically a Storage technology
2)	Magnetic storage	2)	Optical storage
3)	Share the databases	3)	Own/ lease the database
4)	Pay as you use	4)	Pay before you use
5)	More you use more you pay	5)	More you use, less you pay
6)	Prepared searching	6)	Casual searching
7)	Fast searching	7)	Slow searching
8)	Multiple file searching	8)	Single file searching
9)	Fast and real time updating	9)	Slow updating (only by Producer)
10)	Super/Mainframe/Mini systems	10)	PC-based systems
11)	Search, get and pay	11)	Pay, get and search

13.3 MAJOR ON-LINE AND CD-ROM DATABASE SERVICES

CD-ROM medium has been unadventurously used by many publishers simply to distribute their established print or online products in another form without added value in terms of content.

13.3.1 On-Line Databases on CD-ROM Formats

The major databases available on CD-ROM forms are -

Chemical Abstracts on CD-ROM

Chemical Abstracts provides informative summaries of world-wide scientific literature in chemistry and allied subject areas. The CD contains complete CA bibliographic citation with associated abstracts, structure diagrams included with abstracts, issue-wise keyword indexes, CA volume indexes, etc.

Updation: Monthly, including indexes (a total of four to five discs)

Access: Begin with any access point to find the needed information.

Searching: For Bibliographic information: title (words), author name (complete or truncated), organization, patent number.

For Abstract/text: keywords or subject terms

Other access points: Journal title, publication year, document type-journal, printed book, conference proceedings, theses, patents; CAS Registry Number, Chemical structure (displayable)

Indexes: Chemical Substance indexing, Molecular formula (element information), Chemical name (CA index name and synonym, name segments)

Networking: Networking options and dial-in access are available.

Biological Abstracts on CD

Updation: Quarterly

Records: 275,000 records per year

Search Techniques: Free text, Directly from Index, Later Searching, Author Index, Truncation

Subject Searching: Descriptors, Concept codes, Bio-systematic codes, Super taxa

Silver Platter Search Basics: Commands: FIND, SHOW, PRINT
Search Operations: OR, AND, WITH, NEAR, IN

Current Contents on CD

Current Contents is available in four multidisciplinary editions: Life Sciences; Clinical Medicine; Physical, Chemical & Earth Sciences; Agricultural, Biology & Environmental Sciences. It covers 4500 journlas.

Updation: Weekly updates

Searching: Current or Retrospective

Complete Bibliographic data: 52-week rolling file, Reusable search profiles, Searchable author abstracts; User-friendly OVID software, Library holdings tagging facility, Convenient full-text order options

Sci/Tech Reference Plus

* Published by Bowker-Saur.
* Incorporates both bibliographic and directory information
* 135000 Science and Technology books from *Book In Print* database
* 38000 Science and Technical serials from *Ulrich International Periodicals Directory*
* 125000 scientists and engineers from American men and women of science with full biographical details
* 30000 leading US technology firms and subsidiaries

ADONIS - Full-Text Biomedical Journals

* Contains full-text of 600+ major biomedical journals
* Each weekly disc contains full-text of recently published issues of current journals

Inside Information

* Issued from British Library Document Supplied Centre (BLDSC)
* Contains the Table of Contents of 10000 most widely read journals from all fields
* Facility to order reprints electronically through Internet.

13.3.2 Electronic Journals

An.Electronic Journal (or e-journal) is one which is available in electronic form and can be accessed using computer and communication technologies. Depending on the format or source of availability these journals are often referred to as paperless journals, on-line journals, virtual journals, Internet journals, networked journals, CD-ROM journals, etc. These journals are available only in electronic versions, or print as well as electronic formats. The e-journals are found most appropriate by, both publishers and users and useful in getting instant access to scholarly literature. It is estimated that more than 2000 journals are available now in electronic format.

Formats of E-Journals

There is no one well accepted or standard format used in publishing the e-journals. The following are the generally available formats:

1) ***ASCII (American Standard Code for Information Interchange) Format:*** This is a simple text and no formatting or graphics. Eg: *PACS Review, ALCTS News*

2) ***Scanned Images Pages*** (Facsimile) ***using OCR technology.*** Eg: Journals published under TULIP, ADONIS and Red Sage Projects.

3) ***Structured Text Format :*** This format is practically synonymous with SGML (Structured Generalised Markup Language) and HTML (Hypertext Markup Language).

4) ***Mixed Formats:*** Using more than one format, such as SGML, HTML, TeX, PDF (Portable Document Format), PostScript, etc., is referred to Mixed Format.

Electronic journals have several advantages over their counterparts in print format in speed of publication and providing faster access to current literature. These journals also provide facilities for copying/downloading and printing the appropriate articles. Depending upon the license policy of the publisher, they provide multiple access through local networks. These are cost-effective means of publishing as well as acquiring the current literature.

13.3.3 CD-ROM Publishers

With the increased support for CD-ROM products, the number of publishers involved in CD-ROM products have outgrown over the last ten years. According to estimates, there are over 2000 CD-ROM Publishers/distributors in the industry.

CD-ROM publishers play a similar role to online hosts in that they act on behalf of information providers whose data they publish in CD-ROM format. The information provider will compile and own the data published on the disc, but the publisher will provide the necessary retrieval software and will market and distribute the product at a one-off or more usually, annual charge.

Silver Platter Information, CD Plus, DIALOG OnDisc and Macmillan New Media are well known examples of companies which operate with arrangements of this nature with information providers. Silver Platter and CD Plus publish data on CD-ROM format exclusively while DIALOG is an established online host which has moved into CD-ROM publishing also. Macmillan (formerly, Maxwell Electronic Publishing) is an offshoot of the Maxwell Online host.

How to Choose a CD-ROM Publisher ?

While choosing a CD-ROM publisher we have to consider two types of criteria. There are General and Technical.

General:

Do they publish other databases in the same subject area?
How often is the CD-ROM database updated?
When is the disc made available to you?
How much does the disc cost?
What is the policy on offering replacement disc?
Do they have a technical support help disc?
Do they have a 'track record' in the information industry?
How often do they update their software?
Has this or any other product that they produce been reviewed in the appropriate literature?
Do they offer any discounts on second copies, or other databases in the same 'family'?
What is their policy on copyright ?
Do they offer a 'try-before-you-buy' service?
Will they include a CD-ROM as part of a package deal?

Technical:

Does their retrieval software work on all their databases?

Do they offer DOS, Windows, MacIntosh and Networked versions of software?

What is their networking policy?

Do they offer training in their products and software?

On-Line / Internet

If they offer the same database online, does their software allow you to interrogate both the CD version and the online version in one search ?

Will they allow you to download the data to hard disk for further searching that way?

Do they offer any databases which you can search via the Internet?

Checklist of Features expected from a CD-ROM Retrieval Software

General Aspects : Check whether the retrieval software supplied by the producers of the CD-ROM products comply with the following features.

- A single interface for all of a company' products, (with the exception of different platforms)
- Versions of software for a Macintosh and Windows environment which should compare as closely as possible to a DOS-based version
- Ease of use for novices, which should make as much use as possible of function keys, pull down menus and possibly a novice interface.
- On screen, context sensitive help
- Database specific guides, help screens and documentation
- Tutorials, either on disc or available on floppy discs.

Search Features : Check whether the database can support the following Search Features:

- Boolean operators
- Free-text searching
- Field-specific searching
- Multi-disc searching across the same database, or a variety of databases
- Phrase searching
- Marking records for later printing or downloading
- Wild cards and truncation
- Proximity or adjacency searching
- At least one index, preferably several (e.g.: author, country of publication)
- Thesauri on databases as appropriate
- A large number of search statements held on a search history screen for later re-use in the same session.

Display, Printing and Downloading:

Verify whether the following features of the retrieval software supplied by the producer are acceptable.

- Search results should be either immediately shown on screen or one keystrokes away
- Pull down menus to easily re-configure or manipulate the output of results to screen, disk or printer
- Sort search results (for example, by author, journal title or year of publication)
- Download results to database management packages

Saving and Sorting Search Histories :

- Either to floppy disk or hard disk for later re-use
- The option of naming and describing search histories for easier recall at a later time.

Library Holdings:

- The ability to add messages to certain journal titles to indicate their availability within the library
- Show, print or download held titles only

Installation:

- Easy installation either onto a single user workstation or onto a network
- Change installation options without having to reinstall the software from scratch
- A method of ensuring security within the system so that users cannot shell to DOS
- The selection of defaults for displaying, printing or downloading

13.3.4 Evaluation Criteria for CD-ROM/On-Line Services

The evaluation criteria for CD-ROM products relates to mainly two aspects: Search Capabilities and User-friendliness.

Search Capabilities:

1) Is response time (accessing time) reasonable ?
2) Can processing be interrupted ?
3) Are Boolean operators AND, OR and NOT or their equivalents available ?
4) Can multiple word phrases be searched ?
5) Are truncation, wild card, or stem-searching feature available ?
6) May several operators be used in the same search statement with the user specifying which operators will be performed first ?
7) Can previously created sets of search statements be saved and re-used ?
8) May previously created sets of search statements be purged ?

9) Are field-defined searches possible ?
10) Can searches be limited by language and by date of publication ?
11) Can all indexed fields be searched at the same time ? (global search)
12) Can sets be sorted before display or printing ?
13) Can user print more than one record at a time ?
14) Can user select specific records for display or printing ?
15) Can user customise display/print format or choose from a variety of formats ?
16) Can records be saved to the user's floppy disk for use with file managers or word processing software ?
17) Can searcher transfer the search online for more current information ?

User-Friendliness

1) Is there an introductory screen that identifies the database and time span covered ?
2) Is an online screen tutorial included ?
3) Is the user told where to look for prompts and menus ?
4) Is the user shown how to select menus items or respond to prompts ?
5) Is the meaning of commands and menu items explained on-screen ?
6) Is the user told how to back-up through menu screen and exist individual functions ?
7) Is the user instructed in exiting the database and leaving the system ready for the next person?
8) Is function-specific online help provided ?
9) Is context-specific online help provided ?
10) Are useful error messages ?
11) Are examples of commands displayed ?
12) Are examples of logical search operators displayed ?
13) Can the index be browsed for selection of searched items ?
14) Can users select an item from the index without retyping ?
15) Can users select several items from the index without retyping ?
16) Does the system provide suggestions on improving searching vocabulary ?
17) Is explanation of display options thorough and clear ?
18) Is explanation of print options thorough and clear ?
19) Does the system offer short cuts for experienced searchers ?
20) Can explanations not available on-screen be summarised on a one-page printed crib sheet?
21) Is the documentation (user's manual) easy to understand ?
22) Is the documentation arranged in a logical manner ?
23) Is the documentation well indexed ?
24) Does the documentation include samples of searches ?
25) Does the search language use mnemonics and few keystrokes ?
26) Is there efficient use of function keys ?

13.4 INFORMATION RETRIEVAL ON-LINE SYSTEM MODELS AND PROTOCOLS

13.4.1 On-line Systems Models

Technology made it possible commercial online vendors to overcome two barriers - interface and cost control.

The advantages inherent in the online model are — speed of access, frequency of updating, convenience of searching an entire database and also of conducting a search across several different databases simultaneously.

There are two on-line network models: I) Conventional Online Model (also known as, "Multi-Access Model") and 2) Client/Server Model.

Conventional Model has the data, retrieval software and the user interface all running on the host computer system. The user can log-on to the host computer using either a dumb terminal or a PC with communication software and use the services remotely.

Client/Server Model allows the user interface to reside on the local computer rather than downloadable from the host machine. It requires a communication or information retrieval software (protocol) to interact with the search engine. Adequate information retrieval software being developed and standardised and the producers/publishers of databases need to adopt them.

13.4.2 Information Retrieval Protocols

There have been several attempts to achieve standards with the capability of inter-operability between different database systems. Let us study the prominent ones here.

NSO Standard Z39.50

Z39.50 is a standard developed by US National Information Standards Organisation (NISO). Though it was originally adopted as a standard in 1988, it was revised to comply with the international standards (OSI Search & Retrieve). It is a Linked Systems Protocol for system-to-system communication for retrieval of bibliographic information. This is the most prominent standard in networked services and library systems. Many CD-ROM producers and Online service providers are adopting it for their products. Online Computer Library Centre (OCLC), Research Libraries Group (RLG) and National Library of Canada and also the online host, Mead Data Central use Z39.50 standard interface to the range of their products/services.

NISO standard Z39.50 provides the protocol specifications for implementing a basic client-server model of information retrieval. With the proliferation of databases Z39.50 became a key technology for accessing a wide range of resources through a common user interface across computer networks. Under client-server model, one host (called the client, or the "origin") runs a user interface that communicates with an end-user (a human being). The user interface

translates the user's requests into Z39.50 protocol and passes them across a network to a remote machine (a server or a target system). Software on the server then translates Z39.50 protocol back into the server database queries, executes these queries, and optionally passes interim status reports back to the client during execution. When the query is completed, the server uses Z39.50 to inform the client of the final status and size of the query requests. The client can then employ the protocol to request transfer of part or all of the query results across the network from the server; the server employs the protocol to respond to these requests. (Lynch, C.A.:1991)

The advantage of Z39.50 is that the user need not learn the searching conventions of the many host sites. Extensions to Z39.50 will permit the user to retrieve not only bibliographic information but also primary data.

CD-ROM DXS is a data exchange standard, developed by Silver Platter in 1991. It allows CD-ROM inter-operability.

SFQL (Structured Full-text Query Language) is equivalent of the SQL standards in DBMS applications. It was developed by US Transportation Industry.

13.4.3 Internet Services

The Internet is becoming an indispensable resource for information. It provides access to electronic books and journals, reference works, government publications/material, statistical data, directories, indexes, university calendars and documents, maps, graphic images, library catalogues, etc. These are produced and stored in computer systems of different institutions, government and commercial organisations in different countries. The Internet is compared to a network of super highways or an ocean in which a searcher can literally drown in this vast sea of knowledge. To rescue the searcher many navigational tools have been introduced on the internet. Some of the important search tools are given below:

Tools	*Developed by*	*Features*
Archie	McGill University	
Gopher	University of Minnesota	Menu driven system
WWS	Centre for Nuclear Research, Geneva	Hypertext system
WAIS		Menu driven system
VERONICA	Steve Foster & Fred Berrie University of Nevada	Menu driven system

Many of the bibliographic databases and electronic online journals (references with abstracts or full-text), developed by information service providers are also available as Internet sites. Many of the databases which were accessed through commercial on-line service vendors, such as DIALOG, STN, etc., previously, are now accessed by librarians from the Internet. These are accessed through telecommunication networks and dial-up access mode. For instance, DIALOG's IP address is *dialog.com*. The service requires a password and ID to access the system and charge fee for use of the databases in their purview. (In the next two units we shall discuss about the searching the databases and Internet)

13.5 LET US SUM UP

Let us recapitulate briefly what has been discussed so far in this unit.

* In an information retrieval environment, a database is a collection of information that is organised and searachable (and perhaps also updateable) in the context of one or more specific applications.

* Electronic databases are grouped into five types: Reference, Full-text, Directory-type, Factual and numeric, and Pictoral databases.

* Broadly, there are three generations in the evolution of online information systems: Host-base (first generation), Graphics-based (second generation) and Client-Server based (third generation).

* Publishers like SilverPlatter and CDPlus publish data on CD-Rom format exclusively, while DIALOG is an established online host, now moved into CD-ROM publishing.

* Z39.50 is a standard developed by US National Information Standards Organisation (NISO) with a capacity of inter-operability between different database systems.

* With its enormous capability, Internet became an indispensable source for information. LICs use online and CD-ROM databases side by side for information delivery.

13.6 REFERENCES AND FURTHER READING

ACCESS to electronic information./ edited by M Mahapatra *et al.* Bhubaneshwar: Society for Information Science, 1997.

CD-ROM in libraries: Management issues/ edited by Terry Hansion and Joan Day. London: Bowker-Saur, 1994.

COX, John. *Key guide to information services in online and CD-ROM database searching.* London: Mansell, 1991.

ELLIS, David. *New horizons in information retrieval.* London: LA, 1990.

INTERFACES for information retrieval and online systems: the state-of-the-art / edited by Martin Dillon. New York: Greenwood Press, 1991.

LIBRARIANS on the Internet: Impact on reference services/ edited by Robin Kinder. New York The Haworth Press Inc., 1994.

LYNCH, C.A. "The client-server model in information retrieval". IN *Interfaces for Information retrieval and online systems. Op cit.*

PAO, Miranda Lee. *Concepts of information retrieval.* Englewood, Colo.: Libraries Unlimited Inc., 1989.

13.7 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) Define 'database' . Explain the concept and types of electronic databases with suitable examples.

2) Explain online information retrieval services.

3) Discuss the advantages and disadvantages of CD-ROM format.

4) Describe various criteria for choosing a CD-ROM publisher.

5) Write an essay on online and CD-ROM database services with suitable examples.

II SHORT NOTES

a) Electronic journals

b) NISO Standard Z39.50

c) Client-Server Model

d) Internet services

UNIT - 14 : DATABASE SEARCHING; SEARCH STRATEGY

Structure

14.0 AIMS AND OBJECTIVES

There are many approaches to database searching. A well formulated search strategy is essential for all those attempt literature searches. The present unit aims to discuss a number of aspects related to online catalogue searching and formulating strategy for searching bibliographic databases.

After studying this unit, you should be in a position to

- discuss the basic principles of search strategy
- describe the various types of search strategy
- explain the methods of searching online catalogues
- describe the selection and method of searching bibliographic databases
- explain the needs for multiple database searching.

14.1 INTRODUCTION

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

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

14.2 BUILDING A SEARCH STRATEGY

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

14.2.1 Basic Principles of Search Strategy Formulation

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

* Interview the requester
* Conceptualise the search topic
* Use database vocabulary aids
* Interact with the systems
* User system capabilities

1) Interview the Requester

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

2) Conceptualise the Search Topic

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

3) Use Database Vocabulary Aids

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

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

4) Interact with the System

With the advances in technology, searching the online databases in interactive mode and getting the results/printout online have become simpler and cost-effective. Users get familiar

with the strengths and weaknesses of the systems as well as learn tricks that the vendors do not reveal in their manuals and training sessions. It could be learnt only by interacting with the systems. Users will have various options to browse, broaden the searches, modify, restructure, regroup and resume, etc. in interactive mode.

5) Use System Capabilities

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

14.2.2 Types of Search Strategies

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

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

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

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

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

14.3 SEARCHING ON-LINE CATALOGUES

An Online Catalogue is a bibliographic database, containing records of items (i.e., books, journals, microforms, audio-visual material, etc.) available in a library. The online catalogues are designed to access the information online about the library materials by the users with varying backgrounds, age, subject interests, computer literacy, etc.

14.3.1 Characteristics of Online Catalogs

An online catalogue has the following characteristics:

* It is meant to be used by end-users (as opposed to library and information retrieval experts) with or without training in online searching;

* The databse records are usually in the MARC format or derived from the MARC format;

* The record are brief bibliographic descriptions enriched by a small number of controlled subject descriptors (selected from *LC Subject Headings* or *Sears List*) and a classification number (DDC or LC);

* The items are described in the database are usually not limited to a small topic area, but are diverse in subject matter.

14.3.2 Contents of Online Catalogue Records

Online catalogues contain information about the books, journals and other materials in a library or an information centre. The bibliographic records for books contain minimum information of author(s), title, pagination/volumes, publisher, publication year, series, ISBN, subject descriptors, class number, etc.

The bibliographic records for periodicals in an online catalogue usually describe the periodical as a whole and the individual articles are not usually recorded. The contents of a periodical and the description of the articles are recorded in the reference and full-text databases. The online catalogue records are different from the reference or full-text bibliographic retrieval systems that access abstracts of journal articles and even fulltext of the articles. Such retrieval systems usually provide exhaustive indexing of the content of the journal articles. However, the distinction is not clear as some of the online catalogues also provide access to databases that index journal articles also.

14.3.3 Searching Online Catalogues

Online catalogues are usually menu-driven and designed with little search options to cater to even for novice searchers. Some online catalogues operate with command language mode for more sophisticated searching.

Mode of Searching Online Catalogues:

Searching an online catalogue is usually done in two modes: i) Specific Item Searching, and ii) Subject Searching.

i) *Specific-item searching or Known-item searching:* The user tries to locate a particular item that he knows of from the catalogue database. Online catalogues allow searching their databases through author, title, etc.

ii) *Subject Searching:* The user wants to retrieve any item on a particular topic/subject.

14.3.4 Subject Searching

Subject searching is an important activity on the online catalogues. Studies reveal that over 72 percent of searches are essentially subject related. (Poo and Khoo; 1997)

Knowledge needed for Subject Searching:

To perform effective subject searches user require the following kinds of knowledge:

1) Knowledge of the fields that can be used for subject searching and their characteristics;
2) Knowledge of the Thesaurus system from which subject descriptors are selected by indexers;
3) Knowledge of the search capabilities provided by the online catalogues and how to use them;
4) Knowledge of the subject area;
5) Knowledge of search strategies and when and how to apply them.

Fields for Subject Searching

The main fields in the bibliographic records that contain subject information are -

* Subject fields
* Title fields
* Classification Number fields

1) Searching the Subject Fields

The subject fields contains a subject descriptor or subject heading derived from standard lists of subject headings like *Sears List Subject Headings or Library of Congress Subject Headings List, Medical Subject Headings* (MeSH), *Subject Headings in Engineering*, etc. Some databases use available printed thesauri or thesauri compiled in-house.

Many database systems provide online thesauri. The searcher need to have the knowledge of the structure and use of descriptors from the thesauri. A subject descriptor comprises a main descriptor and optionally a number of modifiers called, subdivisions or subheadings. Thesauri use inter-term relationships in their entries, such as broad terms, narrow terms and related terms, besides terminological hierarchies. The searcher has to consult the thesauri and other information retrieval tools in formulating his search strategy. (See details of Thesauri in Unit-8)

2) Keyword Searching in the Title

An alternative approach to Subject Field Searching is Keyword Searching in the Title Fileds. Keyword searching in the tiles is particularly appropriate when there is no descriptor in the subject fields that matches exactly the concept that the user is interested in. The disadvatage is that a concept may be expressed by several synonymous words or variant word forms. Hence, the keyword seaches in the title field may retrieve many non-relevant records. The recall will be more, however the precision of retrieved documents may be low.

3) Searching Classification Number Fields

In this mode of searching, the users need to have knowledge of the classification system and the notation system used in them. The users have to consult the appropriate class schedule to identify the class number. The class number allows the user to find all related works easily. However, the main disadvantage of using class number fields for searching is that only one class number is assigned to an item, even if a work covers more than one topic. In this the recall is low and the precision may be more.

14.3.5 Problems Users Have with Subject Searching

Generally, users face the following problems in subject searching:

1) They have problems in matching their terms with those used in the online cataloguing;

2) They have difficulty in identifying terms that are broader or narrower than their topic of interest;

3) They lack understanding of the printed *Library of Congress Subject Headings* (mainly the abbreviations and subdivisions in LCSH);

4) They don't know how to increase the research result when too little or nothing is retrieved

5) They don't know how to reduce search result when too much is retrieved;

6) They don't know how to use the Boolean Operators and truncation and how to limit keyword searches to specific fields. They are generally not aware of the more sophisticated capabilities of the OPACs.

14.3.6 Improving the Design of the Online Catalogues

The following measures will help in improving the design of the online catalogues:

* Design more helpful interfaces
* Provide non-Boolean "best match" search capability
* Use an automated sequence of search strategies
* Use knowledge-based and natural language processing for query formulation
* Use a hypertext or an enhanced Thesaurus system
* Enhance catalogue records with more subject information
* Enhance the online catalogue for searching class numbers
* Build an expert system front-end to the catalogue system.

14.4 SEARCHING BIBLIOGRAPHIC DATABASES

You have learnt about formulating a search strategy and its importance in successful information retrieval in Section 14.2. In the process of formulating search strategy, the searcher tries to understand the actual information need expressed in possible terms by the information requester and their application in the search process. He limits or modifies the terms to the context, by checking the synonyms, alternative spellings, incorporating scientific and technical

terms as per the requirement. Pertinent journals in the subject field, papers or authors working in the field of specialisation, etc are also helpful in formulating a successful search strategy. Once the search strategy is formulated, it is necessary to select the appropriate database(s) as per the information requirement.

14.4.1 Database Selection

There are a number of databases available and some may be directly relavent to the subject area. More than one database need to be searched in order to get a true picture of subject literature. The subject literature is scattered in a wide range of databases in addition to those at the core.

According to John R Luedtke (1982), there are a number of factors that influence the selection of databases. Some of them are:

1) the type of information desired, which can be bibliographic or non-bibliographic and can range from books, patents or current research to popular articles or statistics;

2) suitable subject coverage and popular degree of comprehensiveness required;

3) the timeliness or time lapsed between publication and appearance in a computerised database;

4) the number of years covered;

5) availability of appropriate indexes to adequately search the requested topic;

6) accessibility of the complete document; and

7) costs of the respective databases.

All the above parameters may not be equally applicable in all situations. Most searches emphasise three to four of them.

14.4.2 Database Selection Method

To select the best database(s) for a specific subject area, it is appropriate to determine the type of literature needed by the requester. When a specialised document or literature type is needed, the requester has to make appropriate selection to list the databases related to that area. The searcher's familiarity and experience with database searching plays a significant role, besides the print aids, provided by the database producers and service vendors, indicating the journal or document coverage. The crossfile search index files of online vendors are useful tools for ranking the subject coverage of databases and for setting up for crossfile searches.

The searcher can scan a selected number of databases for a specific query and obtain numeric results, which help to identify the databases most likely to be useful. It is also better to search the databases other than those routinely searched. (See Flow chart)

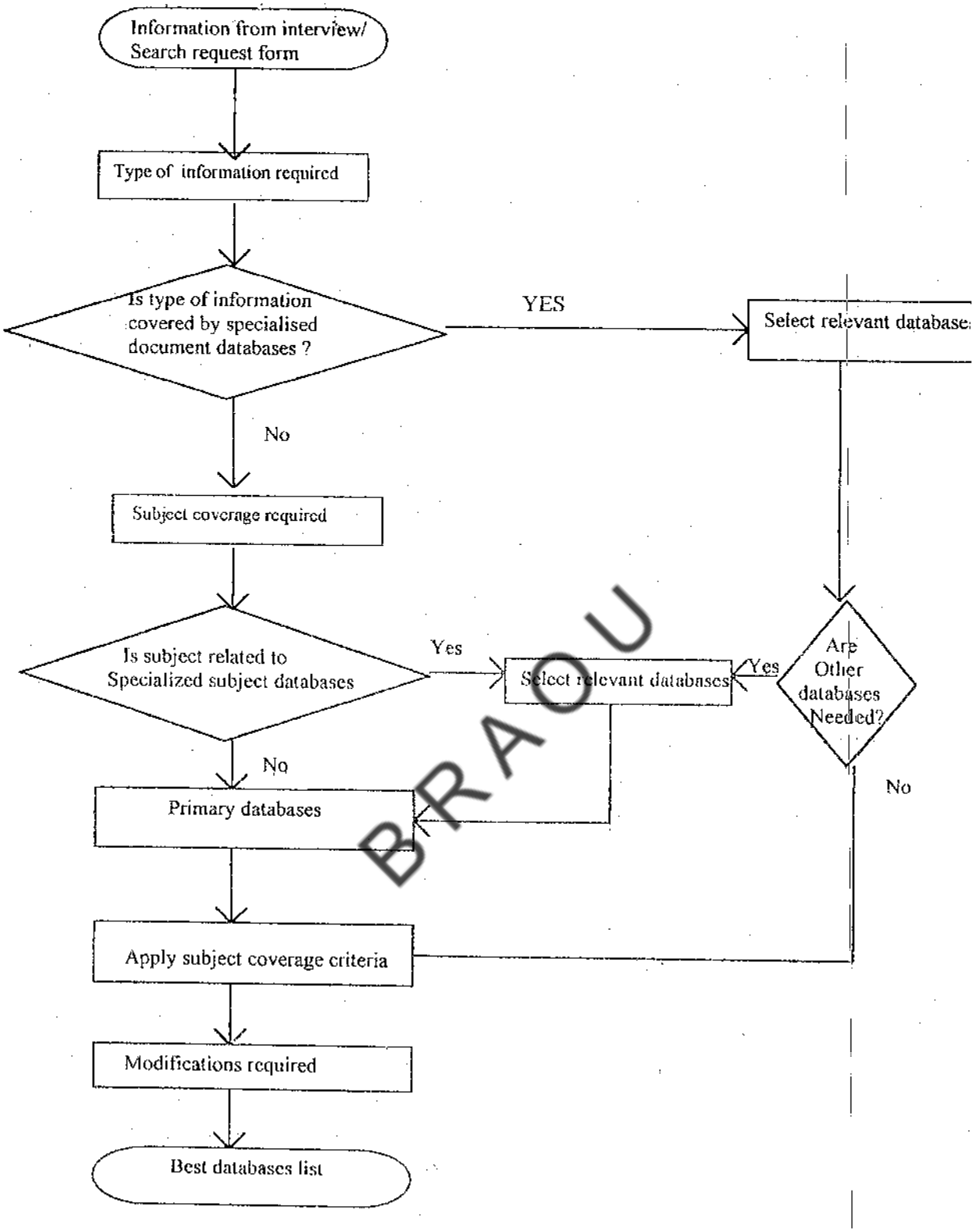


Figure 1. A Database Selection Methodology

Source: *Online Search Strategies*/edited by Ryan E.Hoover.White Plains,NY: Knowledge Industry Publications,1982. P.114

14.5 MULTIPLE DATABASE SEARCHING

Searching one database is not adequate to satisfy the information requirements of the users. In many cases a multiple database search strategy ought to be planned.

14.5.1 Need for Multiple Database Searching

More than one service will need to be searched in order to get a comprehensive literature search. A wide range of sources outside those at the core may contain many relevant documents. The users of abstracting and indexing services and their database equivalents experience certain frustration in searching for literature and information retrieval online and CD-ROM databases. The following reasons may be attributed:

* No single service can be guaranteed to offer full coverage of the literature
* Scope for unique references in each service
* Changes in journal coverage
* Indexed in more than one journal
* Some journals are indexed cover-to-cover, while coverage is selective in some
* Variation in the speed of journals added to databases
* Slow to respond to the appearance of new journals

14.5.2 Multiple Database Search Strategy

The search strategy need to be modified or expanded to suit the database variations. For example, some databases use controlled vocabulary, while others allow free-text searching. Some databases have a policy of segmenting words while others do not. Numeric characters are sometimes converted into alphabets or vice-versa. Search fields may also vary. The manuals supplied by the database producers/ vendors provide search aids with much useful information.

In formulating a multiple database search strategy, it is better to follow worksheet approach suggested by the database producers. Most searches rely heavily on free-text searching, use of truncation and less restrictive logic. Experienced searchers suggest that the most efficient and effective technique is to use all possible forms and variations of keywords (controlled as well as uncontrolled vocabulary) with a high recall logic. If the output is more, then the strategy may be refined. The reverse is not advisable. It is always recommended to use 'Search-Save' capability to save time, energy and money by not having to re-renter the strategy for subsequent search runs.

14.5.3 Searching Indexing and Abstracting Services

The indexing and abstracting services in print or database form are most useful for conducting literature search. These services offer access to a wider range of source documents, especially journals, conference proceedings, theses, reports, etc. Generally, the abstracting and

indexing services cover journal literature heavily and a fair degree of coverage is usually given to conference proceedings and other forms.

Though both indexing and abstracting services are key to literature scattered among major journals and other primary sources, the basic difference is that indexing services provide indexing, but do not,summarise the subject content of the references which they list. The provision for summary/abstract of the article will add value to the list of items provided by the abstracting services. The publications typically appear on monthly basis and their indexes are cumulated annually. The annual cumulations help in retrospective searching. Advantage of online databases is that their cumulations are uptodate.

Searching of online/CD-ROM databases of indexing and abstracting services depends on how best the searcher identifies the relevant subject sections in order to make full use of the subject index provided. References in all abstracting and indexing services can be located by author's name, while additional access points provided may include journal source and author's affiliation.

14.5.4 Searching Bibliographic Databases in LIS

A number of indexing and abstracting services are availble online and CD-ROM forms in a subject field like their print equivalents. For example, in library and information science, we have *Library Literature, Library and Information Science Abstracts and Information Science Abstracts.* Though there is some overlapping/duplication of literature, the scope and coverage varies in these sources. Any search for current literature in LIS need to be to conducted on these three secondary sources of information. Let us discuss the nature, coverage and organisation of information in these three sources.

S.No.	Title	Frequency	Indexes	Access
1)	*Library and Information Science Abstracts* (LISA)	Monthly	Subject, Name	Online CD-ROM
2)	*Library Literature*	Bimonthly	Subject, Name (in one sequence)	Online CD-ROM
3)	*Information Science Abstracts*	Monthly	Subject, Author	Online CD-ROM

1) Library and Information Science Abstracts (LISA)

The *Library and Information Science Abstracts* (LISA) is one of the most important secondary information source periodical publications in the field of Library and Information Science. LISA is compiled by the Library Association (LA). It began its publication in 1950 and monthly issues started in 1982. It covers around 350 journals in the field of library and information science. Online and CD-ROM versions are also available.

The entries in LISA contain bibliographic details with informative abstracts. The entries are arranged according to a faceted classification, which uses alpha-numeric notation. The dominant subject aspect of the document determines its position in the classification and hence users need to search under it. Instead of relying on the most obvious section of the classification scheme, the searcher should use the subject index to group related items which are scattered by classification. The other possible search route is through the name index. Besides author names, this index contains the database and their host names. This makes it easy to locate references about a particular file or the host through which it is available.

2) *Library Literature: An Index to Library and Information Science.*

Library Literature is published by H.W. Wilson Company, New York since 1934 (Bimonthly issues from 1969). It covers over 220 journals. The printed version is arranged in a single alphabetical listing in which entries are grouped under headings for subjects, authors or proper names such as databases.

Library Literature is available online and CD-ROM forms. Subject access is served by headings of varying specificity. The subject headings also use sub-headings (e.g.:Evaluation, Reviews, etc.) to specify the content. 'See' and 'See also' references are also used.

Library Literature is available online and CD-ROM formats with retrospective coverage extends only to 1984.

Though the source coverage in Library Literature is narrower than that of LISA, it scores well on two counts. First, it aims to index its journals fully and to provide access to items, such as conference reports, trade exhibitions, review of new publications, etc. Secondly, the time lag between publication and indexing of entries appears to be kept within respectable bounds.

3) *Information Science Abstracts*

It first published in 1966 under the title Documentation Abstracts and changed to the present title in 1969. Since 1984, it has been issued as a monthly publication. It indexes more than 400 journals and the coverage is international. It also includes computer science literature, besides librarianship and information science. It contains a high proportion of abstracts for report literature. Books are sometimes listed, with abstracts for individual chapters if appropriate.

Entries include full bibliographic details and informative abstracts. Arrangement is by a classification system of eight main sections and sub-sections, representing the Information Science. Cross references are provided at the end of sections to guide the searcher to relevant entries in other sections. The subject index aids the searchers to search for specific topics by providing a mixture of free-text and controlled vocabulary terms for retrieval. Country names and general subject categories (e.g.: Legal Information, Medical Information) may also assist retrieval. Host and database names are included in the subject index on a selective basis. Both indexes are cumulated annually. The cumulated subject index only gives the accession number for references listed under a heading.

14.5.5 Sample Search

John Cox (1991) made an interesting study on how the above discussed three secondary sources of information in LIS treated the topic 'On-line Information Retrieval' and 'CD-ROM Searching' in their databases.

LISA groups the literature on the above topic separately at:

Zm Onlihe Informatin Retrieval; and
Zjjc CD-ROMs.

The *Library Literature* treats the topic 'On-line Searching' as general heading underneath which cross-references are provided to guide the searcher on related topics, often under more specific terms such as 'End-User Searching', 'Information Systems' and 'Online User Groups. There are no cross-references to the heading CD-ROMs, though a separate section for CD-ROMs is present.

Information Science Abstracts provides main sections 'Information Systems and Approaches' and 'Information Storage and Retrieval'. 'Bibliographic Search Services, Databases' is the sub-section under the first one, while CD-ROMs is covered under the sub-section 'Storage' in the main section of the later one.

Source List of CD-ROM Database Search

Topic : CD-ROM + Information Retrieval
Database : ***Library and Information Science Abstracts*** (LISA) 1999

CD-ROM - an information stor	Sharma, R. S.		LS 1998
CD-ROM: a new advance in med	Imam, S. S.	Pakistan Library	LS 1996
Moznosti vyuzivania technolo	Pracna, M.	Kniznice a Inform	LS 1995
Information retrieval from c	Rutteus, B.	Audiovisual Libra	LS 1995
A comparison of information	Large, A.	Information Proce	LS 1994
An investigation of children	Perzylo, L.	Australian Librar	LS 1994
Comparison of some widesprea	Valas, G.	Online and CD-ROM	LS 1994
A comparative analysis of in	Haner, B. E.		LS 1993
CD-ROM interfaces for inform	Shaw, D.		LS 1993
An investigation of children	Perzylo, L.	Microcomputers fo	LS 1992
CD-ROM in information storag	Shah, G. A.		LS 1992
Vso do CD-ROM na recuperacao	Dias de Andra	Ciencia da Inform	LS 1990
Bewertung von Alternativen i	Lehmler, Wilf	Bibliotheksdienst	LS 1989
Effectiveness of information	van der Walt,		LS 1989
Information retrieval skills	Aguado, Patri		LS 1989
CD-ROM e recupero dell'infor	Russo, Robert	Bibliotecario	LS 1988
CD-ROM or online for medical	Morgan, V. El		LS 1988
Information retrieval on a m	Sieverts, Eri		LS 1988
Perspectives on...CD-ROM for	Lunin, Lois F	Journal of the Am	LS 1988
Information retrieval from C	Borgman, Chri	Canadian Journal	LS 1987
Information retrieval-databa	Ekengren, Bo		LS 1987
CD ROM technology: a new era	Herther, Nanc	Online	LS 1985
The Status information stora	Verscheor, C.		LS 1985
Applications of CD ROM in in	Nicholson, D.	CRLIS CR	

A Sample Record from LISA Database

Library and Information Science Abstracts
79283

CD-ROM - an information storage and retrieval tool.
R. S. Sharma
R. Patnaik
Information management in academic and research libraries. Proceedings of the 5th National Convention for Automation of Libraries in Education and Research (CALIBER-98), Bhubaneswar, India, 4-5 March 1998. Edited by M. Mahapatra et al. Ahmedabad, India: INFLIBNET Centre, ISBN 81-900825-1-5, 1998, p.65-68.

Paper presented at the 5th National Convention for Automation of Libraries in Education and Research (CALIBER-98), at Bhubaneswar, 4-5 March 1998. CD-ROM databases are cost-effective. Due to advancement of electronic sector more and more users are using CD-ROM for retrieving information as it follows Hypertext system. They are unique because large volumes of data can be stored at a very low price. The information contained on CD-ROM includes text, still images, audio, digital video and animation. Describes the advantages of CD-ROM and the status of the CD-ROM server at SAC Library, ISRO. (Original abstract - amended)

English
1998

14.6 LET US SUM UP

Let us recapitulate briefly what has been discussed so far in this unit.

* Search Strategy encompasses all activities involved with searching, from the reference interview through the verification of search results.

* The basic principles of an effective search strategy involve - interview with the requester, conceptualisation of the search topic, using database vocabulary aids, interaction with the systems and understanding the system capabilities.

* Search strategies can be divided into Initial Strategies and Reformation Strategies. Reformation strategies are used to improve the search results after reviewing the result of the initial search.

* An online catalogue is a bibliographic database, containing records of items, ie., books, journals, microforms, audio-visual material, etc. available in a library.

* Subject searching is an important activity on the online catalogues. Studies reveal that over 72 per cent of the searches are essentially subject related.
* In selecting a bibliographic database for searching, factors such as type of information

covered, comprehensiveness, timeliness, availability of indexes, costs, etc. should be taken into consideration.

* No single database service can be guaranteed to offer full coverage of the literature, hence, there is a need multiple database searching.

14.7 REFERENCES AND FURTHER READING

COX, John. *Key guide to information services in online and CD-ROM database searching.* London: Mansell, 1991.

HOOVER, Ryan E. "Online systems, yesterday and today". IN *Online Search Strategies/* edited by Ryan E. Hoover. White Plains,NY: Knowledge Industry Publications, 1982.

POO, Danny CC and Christopher Khoo. "Subject searching in online catalog systems". IN *Encyclopedia of library and information science/* edited by Allen Kent and Carolyn M. Hall. New York: Marcel Dekker, 1997. Volume 60; p.324-340

14.8 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) Explain the basic principles and the types of search strategy.

2) Describe the characteristics of online catalogues and discuss the methods and problems of searching them.

3) How do you select a database for searching bibliographic information ?

4) Explain the various reasons for multiple database searching and discuss the search strategy to adopted.

II SHORT NOTES

a) Vocabulary aids

b) Interview with the Requester

c) Free-text searching

UNIT - 15 : SEARCHING THE INTERNET

Structure

15.0 AIMS AND OBJECTIVES

The Internet is a network of thousands of computers scattered across the globe that allows free exchange of information. The present unit aims to discuss the role of Internet in information retrieval.

After studying the unit, you should be in a position to

- describe a brief history of the Internet
- explain the common information services available through Internet
- discuss the impact of Internet on information retrieval services.

15.1 INTRODUCTION

Today the Internet has become a buzz word of the computer world as well as library and information profession. Internet is fast becoming an important channel for communication. Ten years ago, few would have predicted the amazing growth and universal acceptance of Internet for sharing and exchanging information. Now every one would agree that Internet has revolutionised the human communication. Users know how quickly the Internet grows and are eager to use its resources and services. Millions of users send messages, listen to music, participate in discussion groups, engage themselves in chat, read magazines and newspapers from across the world, study the scholarly journals, and watch video. The Internet bridges the distance, time and cultures and brings all communities into its fold. The Internet is a network of thousands of computers scattered across the globe that allows free exchange of information. The knowledge of Internet resources is becoming as a basic knowledge to librarians as well as general public.

Internet has created a major impact on library and information services throughout the world than any system, technique or technology worth comparable. The knowledge of Internet resources is becoming as basic as knowledge of an index, a classification scheme, an online catalogue, or any other tool the library and information professionals use to serve their clientele.

In Course-05 Application of Information Technology, we discussed the concept, tools and services of the Internet (see Unit-16: E-Mail and Internet). The present unit provides much emphasis on information retrieval aspects of the Internet.

15.2 INTERNET - A BRIEF HISTORY

The Internet was started in 1969 as defense project, namely, ARPANET, in the USA. It was developed by US Department of Defence's (DoD) Advanced Research Project Agency (ARPA). Its main objective was to ensure uninterrupted flow of information during war period. Later, it was developed in the universities of USA with few computer networks for educational and research purposes.

In the first decade it was used to facilitate e-mail, support online discussion groups, allow access to distance databases and support the transfer of files between government agencies, companies and universities.

During the early 1980s, all the interconnected research networks converted to the TCP/IP protocol and started exchange of information among educational institutions in and outside the USA. The ARPANET became the backbone of the new International network. Thus, the Internet was born.

In 1990, a hypertext-based Internet protocol was introduced to facilitate the graphic information. With this enhancement the Internet became a virtual hypertext network called the World Wide Web (WWW). It was informally renamed as the Web.

Now, the Internet is a global network of commercial and non-commercial networks. Interestingly, commercial information has been dominating when compared to the other sectors.

Internet is decentralised and there is no regular maintenance body, no central agency or institution that sets rules. Instead, users of the Internet formed a society known as Internet Society to promote information exchange and development of the Internet technologies.

15.3 INTERNET CONNECTION

Internet is a network of computer networks spread all over the world. It is a series of networks linked together and uses very precise rules to communicate with each other. It allows any user to connect to and use any available network or computer on the Internet.

15.3.1 Internet Protocols

Internet uses physical connections usually telephone lines, direct wires, fibre optics, and satellite transmissions to link one computer network to another. All the computers on the Internet will 'speak' to each other through a communication standard format known *Transmission Control Protocol/ Internet Protocol* (TCP/IP). This protocol breaks the data into small pockets with an address label on each pocket and sends them to their destinations through different possible routes. There, all the pockets are reassembled into original form. This is called Pocket Switching mechanism. This Pocket Switching mechanism will ensure the data transmission to the destinations through different routes. This uninterrupted flow of data is possible if the whole communication channels are completely devoted. Many organisations use leased telephone lines.

15.3.2 Internet Service Provider (ISP)

Usually, connectivity to the Internet is provided by a company, called Internet Service Provider (ISP). The company provides its customers with access to the Internet, typically through Dial-Up Networking. Customers pay monthly or annual fee to the ISP. In India, Videsh Sanchar Nigam Limited (VSNL) is the major Internet service provider. The National Informatics Centre (NIC) and Education and Research Network of the Department of

Electronics (Government of India) provide Internet connection to the government, educational and research organisations respectively. But due to the Government's liberal policy, private ISPs have been now entering into the business of providing the Internet connectivity. Access speed, customer per modem ratio, number of e-mail accounts provided for each user, storage space provided online and type of licensed internet software are some of the important aspects to be examined before deciding the ISP.

15.3.3 Internet Connection

In order to get an Internet connection at office or home, you need to have a computer (with Windows operating system with a Browser) fitted with a modem, a telephone line and shell or TCP/IP subscription from any ISP. Microsoft Windows 95/98 and Windows NT have built-in support that enables us to connect to an ISP. A Point-to-Point Protocol (PPP) or Serial Line Internet Protocol (SLIP) account with the ISP is needed. The ISP may require the user name, password, local access phone number, etc. This information has to be entered into the system's Dial-up Networking configuration to connect to the Internet successfully. Usually, the ISP assigns the host and domain name, and DNS server and IP address.

Domain Name System (the DNS) is an standard addressing system on Internet. It is similar to our home address. Every computer on the Internet will have a unique address. This is called IP address and it is usually a number. Since it is difficult to remember numbers, the DNS provides meaningful and easy to remember names to all computers. For example, vsnl.net.in, hotmail.com, andhrapradesh.com, thehindu.com, worldbank.org, whitehouse.gov.

15.3.4 Internet Browsers

Browsers are programmes that act as an interface between the user and the Internet. They enable the users to read hypertext on files on the Internet. Basically, there are two types of browsers - Text-based Browsers and Graphic User Interface (GUI) Browsers. The most popular browsers, Netscape Navigator and MS-Internet Explorer, come under GUI. Systems and Systems that lack graphics can use LYNX, a text-based browser.

Internet Explorer (IE) comes along with the Microsoft software (Eg: Windows 95/98). In fact, the Microsoft offers IE free to public and vowed that it will always be free. The IE could be installed (or setup) along with Windows.

You may start your internet browser in the same way you start any other applications. A click or double-click its icon on the desktop, or select it from the menu. You may also start the browser after connecting the dial-up networking to your local server. Once you start your browser, it automatically loads a starting page (home page of the browser) by default. Type the address of the desired page into the Address or Locations textbox near the top of the browser window.

15.3.5 Browsing

The Browsers are used for getting information from the Internet. Browsing means meandering through available electronic information or skimming on information resources or looking simply for interesting items. The metaphor corresponds with paying a desultory

visit to your favourite bookshop or a library to peruse the books. In the context of Internet, it means going through specific sequence of hyperlinks that a particular user selects and moving through a large hypertext or hypermedia to get required information.

Surfing and Navigating are the two other terms used with the Internet for getting information. ***Surfing*** is the practice of browsing through the contents of newsgroups, web or other information services on the Internet. Like a surfer riding one wave and then another, the user on the Internet visits the sites one after the other. ***Navigation*** means to sail over the Internet sites for information. These two terms, Surfing and Navigation are used to bring in symbolism between the sea and the Internet.

Every page on the web has an address, called ***Uniform Resource Locator*** (URL). A URL (most often pronounced as "Earl") is the standard address that indicates the location of a page or resource on the Internet universally. The location can be a computer of a company (ending with ".com"), organisation (.org), government (.gov) or educational institution (.edu), etc. A URL is often referred to as a Page Address. Computer magazines *(Eg: The Net, Net Guide, Internet World)* and Internet newsgroups are good sources for finding interesting web addresses to visit. A typical URL may look like *http://www.microsoft.com/.*

15.4 COMMON INTERNET SERVICES

Today, when you say 'Internet', you really mean you are talking about e-mail or world wide web (www). But in reality, Internet is more than e-mail and www. Its meaning is much wider. Internet encompasses a number of services, which could be grouped into four types:

1) Information Retrieval Services (FTP and Gopher)
2) Information Search Services (WAIS, Archie, Veronica)
3) Communication Services (E-mail, Telnet, UseNet, IRC)
4) Multimedia Information Services (World Wide Web)

15.4.1 Information Retrieval Services (FTP and Gopher)

There are two popular information retrieval services, namely File Transfer Protocol (FTP) and Gopher.

i) File Transfer Protocol (FTP)

The most common service used for sending files between computers on Internet is File Transfer Protocol or FTP. These files can be data files or programmes files. With FTP you can have access to millions of computer files. FTP will let you download or upload transfer any type of files — programmes (including software and games), text, pictures, sound or any other file format. FTP allows mass distribution of documents created by various universities, documents held by libraries, free software or shareware, maps, satellite photos, etc.

In the case of text files, the procedure is simple. However, transferring pictures and software is a bit complex.

When you access the FTP programme, it asks for your name and password. If you don't have, anonymous FTP retrieves files that are available to the general public. In that case the procedure is to give anonymous as user name and then type your e-mail address in the place of the password. For example, the address of an FTP will be ftp://ftp.cdrom.com/

ii) Gopher

Gopher is similar to FTP. It is a menu-based programme that enables you to browse for information without having to know where the material is specifically located. One of the nice features of a gopher is that it is very easy to browse and retrieve files from the Internet than using FTP. Once you enter the Internet through a gopher, you simply follow a set of menus or directories to browse for information. Many foreign universities use a Telnet or Gopher interface for their offerings. Today, we can access catalogues of several libraries such as Library of Congress, all national libraries in USA and Europe. In India, libraries of IITs, and a few university libraries offer their catalogues through Telnet.

Gopher is supported by the University of Minnesota. Gopher is gradually being superseded by http and browsers such as Netscape, which can reach Gopher sites as well as http information.

15.4.2 Information Search Services (WAIS, Archie, Veronica)

i) Wide Area Information Service (WAIS)

WAIS is a system that searches for your subject through documents on all servers all over the world. WAIS searches a set of databases that has been indexed with keywords, and returns addresses where you can locate documents that would be of interest to you.

The key features of the WAIS system are - (a) The WAIS system has the ability to have indexes that actually point to other WAIS services, and (b) the heart of WAIS system is the use of client software running on your local computer that lets you ask for information in simple, English-like language. To use WAIS, you have to telnet to a WAIS server, or locate WAIS-based resources on a gopher menu.

ii) Archie

Archie was the first of the information retrieval systems developed on the Internet. The Archie database system was developed by the Alan Emtage, Bill Heelan and Peter Deutsch at McGill University.

Archie is used for finding resources that are available on Internet. Archie creates a central index of files that are available on anonymous FTP sites. We can have Archie client running on the local machine or use Telnet to connect to one of the Archie servers and search the world-wide network of databases to find a file.

iii) Veronica

Veronica (Very Easy Rodent-Oriented Net-wide Index to Computerised Archives) is a service that searches menu items on Gopher servers. To use Veronica you have to be connected to a Gopher server that gives you access to a Veronica server. Because Gopher menu items can be descriptive phrases (more than just file names), it can be easier to find information of Internet through Veronica than it is through Archie. Veronica may find a file on an FTP site that Archie would not, because you can use Veronica to search for information on topics (maps, for example), rather than just searching for file names. Veronica allows the searchers to use either a single keyword or a string constructed of an unlimited number of ANDs and ORs (Boolean operators) with keywords.

When Veronica has finished searching Gopherspace it builds a Gopher menu that contains all of the items it has found to match your search. You can then examine those items by selecting them, just as you would from any Gopher menu.

15.4.3 Communication Services (E-mail, Telnet, UseNet, IRC)

i) Electronic Mail

Electronic Mail or e-mail is considered to be the fundamental service of the Internet. E-mail is the today's popular communication medium through which you can send and receive messages. Based on the e-mail technology, Internet offers a variety of services. These include listserv or Electronics Discussion groups, Bulletin Boards, Newsgroups. Through these services, you can send mail to any number of individuals, participate in discussion and pose queries to a group of individuals. It is very cheap and fastest mode of sending messages across the globe. Through e-mail, now you can send drawings, photograph images as attachments at almost no cost with the better quality than a fax.

E-mail addresses vary from site to site in their format. It contains a personal identification mark to the individual, a site address or machine name, and an identification of whether it is a commercial site, an educational site, or an organisation. For example, the e-mail address of your university is *braouap@hd1.vsnal.net.in.* and the personal e-mail address of the writer of this course unit is *vcsrao@hd1.vsnl.net.in.* The first part identifies the organisation's/personal name, while the second part hd1.vsnl.net refers to the VSNL Server1 of Internet located at Hyderabad and the last one identifies the country (India). Users of the Internet can also subscribe to free e-mail facility, such as *hotmail.com, coolmail.com* and *yahoomail.com.* Except the telephone connecting charges, the users can send their e-mail at free of cost.

Listserv are electronic groups that typically center around a broad topic (for example, Internet and Libraries). Hundreds or thousands of people (may be of a profession or association) having access to e-mail facility join together and form a listserv. It does not cost anything to subscribe to a listserv. Every e-mail message sent to the listserv is distributed to all the members of the listserv group. The user of the listserv sends an e-mail message with a request: *subscribe [listserv] Firstname Lastname*. Users can also use requests unsubscribe, search the archives, etc.

In India, IASLIC has recently started "IASLIC-LIST", a global LISTSERVER, to facilitate fast and economic communication between the IASLIC and its members and also enable IASLIC members to communicate amongst themselves. To join or subscribe to this list, you can send an e-mail ***SERVER@LISTSERV.INDIAX.COM*** with the command JOIN IASLIC-LIST.

Bulletin Board Systems (BBSs) connected to the Internet are very effective in sharing the information online. Users can access the electronic BBSs without the need for long distance telephone call. The BBSs became a means for the sponsor to issue announcements, distribute software and communicate personally with the users, etc.

ii) Telnet

Telnet allows you to use any host computer on the Internet as if you were sitting in front of it and directly connected. It provides direct path so that remote computer communicates directly with the terminal you are actually using. Therefore, Telnet is often called as *Remote* Login. The user can enter data, run programmes, or do any other operations. Usually, you need a password to access and use the computer at the remote location.

iii) Usenet

Usenet is a set of thousands of newsgroups (discussion forums) distributed via Internet. The messages in Usenet are organised into thousands of topical groups or "newsgroups", which cover specific areas of interest. Usenet is read and contributed to on a daily basis by millions of people. A user can read news, ask questions, answer questions and participate in the discussion. Newsgroups have names such as -

news.	for announcements about Usenet itself
comp.	for computer science and technology
sci.	for academic aspects
soc.	for cultural interest groups
rec.	for hobbies and sports
talk.	for wide ranging discussions, often heated
misc.	for a few topics that don't fit elsewhere
alt.	for trial newsgroups and alternative topics.

Usually, the Usenet addresses are in the form of —

psuvax!ugacc!aisun!!mcovingt

It means that the user McOvingt on machine Sun, which can be reached through UGACC, which can be reached through PSUVAX. Because the cost involved in distributing every message is paid by the sites all over the world and not just the sender's site. Therefore, it is extremely important to use newsgroup only for its intended purpose. Except when specifically required, do not use it for advertisements or mass solicitation.

iv) Internet Relay Chat (IRC)

IRC is a protocol and a client-server programme that allows you to chat with people all over the Internet, in channels devoted to different topics. It helps you communicate (chat) with people with similar interests "live" (in real time), so that you need not wait for people reply to your messages.

Using the 'talk' command (talk username@address) you may send words directly to someone's screen. There are three major networks of IRC servers, namely, EFNet, Undernet and DALnet. For example, the page address of IRC server of Undernet site is *http://www.undernet.org/.*

15.4.4 Multimedia Information Services (World Wide Web)

The very popular service on Internet today is World Wide Web. This is also called, in short, as Web or WWW.

The WWW is one of the newest client-server based Internet services. It uses the standard format of display developed by CERN (the European Lab for Particle Physics) in the late 1980s. This format allows anyone easily access and display documents that were stored on a server anywhere on the Internet. This also allows links to other documents to be placed within documents. Soon after the WWW service was made public and it came to extensive use in the Internet.

WWW is a multimedia linked database system that spans the globe. The web combines text, pictures, sound and even animation. The web with its colourful pictures, music and moving images made Internet much more accessible and fun to use it. With a click of mouse button, the browsers (Netscape or Internet Explorer) allow you to navigate the wonderland of web.

The hypertext documents contain commands from a language called, Hypertext Markup Language (HTML). The HTML also enables the display of multimedia files (such as movies and sounds in the documents) by the graphical WWW clients. Some WWW clients also display multimedia player programmes.

Most WWW clients also allow you to access other Internet services, such as FTP and Gopher.

15.5 SEARCHING THE INTERNET

Now let us know about different services on the Internet of which most of them are available via the web.

15.5.1 Web Page and Home Page

Every place on the Internet has an address. This is usually known as Web site and each web site has information in the form of web pages. The web pages are hypertext documents on the web. The page you begin with when you start your browser is called *Home Page.* It is the

main page of a web site. When you start Internet Explorer or Netscape Navigator, its home page comes up automatically (until or unless you change it to start at a different page). You can stop loading the default home page by pressing the Stop button and then you can enter an address of the required web site directly. Hypertext marked links on the home page help you search other pages of the web site. Book marks enable you to jump back to an interesting page.

A web site address (also called United Resources Location, or URL in short) usually begins with http protocol. The web page address has three parts: protocol, host name and file name. The syntax will be

Protocol://hostname/filename.

For example, *http://www.microsoft.com/* means that the page is of Microsoft computer server on the Internet access tool/service, namely web (www), which can be accessed through http (Hypertext Transfer Protocol), a standard format for publishing information on the Internet. The third part is the file name of the specific document to be located according to the query. All the web sites follow the same pattern.

Some of the prominent library related web sites are

American Library Association	http://www.ala.org/
Library Association(UK)	http://www.fdgroup.co.uk/la.html/
Library of Congress catalogue	http://lcweb.loc.gov/catalog/
Encyclopedia Britannica online	http://www.ed.com/
OCLC	http://www.oclc.org/
BlaiseWeb British Library (UK)	http://blaiseweb.li.uk/
IFLANET: IFLA Canada &Netherlands	http://www.nlc-bnc.ca/ifla/http://ifla.inist.fr/
Research Libraries Group(RLG)	http://www.rlg.org/
Directory of Worldwide India related resources	http://www.hindustan.net
Internet bookshop	*http://www.bookshop.co.uk/*

15.5.2 Search Engines and Gateways

If you know the web site address, you can simply type it in the browser's URL and you are instantly there. If you don't know the address, don't worry,- you have a variety of search engines and gateways, they will search the site or information you want to have immediately. Today's popular search engines are *Infoseek, HotBot, Lycos, Altavista, Web Crawler* and many more. They help you to find relevant web sites on the Internet. There are also some meta-search engines which help you find the search engines. The web page addresses of some popular search engines are -

Altavista	http://www.altavista.com/
Infoseek	http://www.2.infoseek.com/
Lycos	http://www.lycos.com/
Hotbot	http://www.hotbot.com/

Mamma	http://www.mamma.com/
Alltheweb	http://www.alltheweb.com/
WebCrawler	http://www,webcrawker.com/
Whowhere	http://www.whowhere.com/
Sharewhere	http://www.sharewhere.com/

A ***Gateway*** is a computer that connects one network to another for the purpose of transfer of files or e-mail messages, when the two networks use different equipment, such as mainframe or micro-computers that do not even share the same routing and protocols (For example, NetWare to TCP/IP). A gateway could also be a computer that transfers posts/files from a newsgroup to a list and vice-versa. For example, if you send a message from your NICNET or America Online (AOL) account to someone at CompuServe, or elsewhere, then you are sending messages over different networks of the Internet through a gateway. *Whois* and *WhoWhere* are the gateways help you to find the e-mail addresses of people at the InterNIC site (the Internet's Network Information Centre).

15.5.3 Searching the Web

There is no single, definitive way to search the web. Because Net is changing so rapidly that any comprehensive listing of sites may become obsolete in no time. Therefore, it is always better to try several different approaches. Generally, there are two models for finding specific information on the web — (i) Searching through a Directory, and (ii) Searching with a Search Engine.

i) Searching through a Directory

One of the best directories on the web is the Yahoo site. Yahoo is organised hierarchically with site resources edited. It starts with a general topic area and narrows down to a more specific topic. For example, the Health topic shows the topics Medicine, Drugs, Diseases, Fitness, etc. If you click on the Medicine link on Health:Medicine Page, it leads to other subtopics such as Anatomy, Cardiology, Medical Schools, etc. If you click on any subtopics, you can go to the page listings' on that subject hierarchically. Some important search directories and their web addresses are given below:

A2Z	http:a2z.lycos.com/
Excite	http://excite.com/
WebDirectory	http://www.webdirectiry.com/
Yahoo	http://yahoo.com/

Thus, you can reach topics of interest to you by more than one route. You can search the homepage of the directories (for example, Yahoo) through keywords. To perform a search type a word or words in the box near the top of the Page and then click on the search button. The search directory quickly returns the list of sites relating in someway or other related to your keyword. The hyperlinks will lead you to interesting documents on the Net.

ii) Searching the Web with a Search Engine

As you know a search engine is a Web site, designed to perform searches on the Internet. Unlike Yahoo, most of the search engines attempt to include every single page on the Web. These entertain even complicated search queries. Some of the favourite search engines are given in the Section 15.5.2.

You can enter your keyword(s) into the box and click on the search button. The search engine will return a list of sites ranked in order of their likeness to match your keywords. The hyperlinked listings lead you to interested documents through which you can surf the Internet.

15.5.4 Searching off the Web

The web is just a part of the Internet and there are many resources that are hidden. You can explore them through search engines and gateways. However, some of the non-web resources can be even flakier than the web when you are trying to access them. Be careful !

i) Searching Usenet

Usenet and related newsgroups are the public discussion bulletin boards available on the Internet. The news articles that are posted on the Usenet may expire after some specific period (may be weeks or months) depending on the news server. *DejaNews* is probably the best search engine to help you search the Usenet for news. When you type word(s) in the 'Search For' box and click on the Search Button, the search engine will display a list of articles containing the word(s) you asked for. The items on display also provide links to the article's contents and author's profile.

ii) Searching Gopherspace

Before WWW was introduced, Gopher was the easiest way to browse the Internet. Gopher presents a list of menu items, from which we have to chose item to access a document or another menu. Gopher sites are not indexed with the major search engines, hence we have to depend on Gopherspace. Gopherspace is the set of all Gopher servers serving the Gopher menus and documents. To search on the web, first point your browser at *gopher://veronica.psi.net/2347/7-t1%20%20.* When you can enter keyword(s) in the text box, the search page will list an ad-hoc Gopher menu with items that match your search words.

iii) Searching University Libraries

Many university libraries on the Internet are available on Telnet or Gopher interface and hence, they are not easily reached via web search engines. However, through Hytelnet programme, you can reach the web for an up-to-date list of university libraries. For this, point your browser at http://library.usask.ca/hytelnet/. To look for a university library, click on Library Catalogs and then choose the geographical region and the university. To perform a general search of university libraries, click on the Search Link at the top of the menu of the Hytelnet Page. Then enter the keyword(s) in text box and press the Search button. Thus, you can have access to OPACs of the university libraries.

iv) Searching the databases

Many of the reference and full-text databases librarians used to access through commerciaal online systems are now available through the Internet. They charge for use of the databases and hence these services require a password and ID to access them. Each service has a telnet address. For example, the telnet address of some of the commercial vendors are

DIALOG	dialog.com
OCLC	epic.prod.oclc.org
STN	stnc.cas.org
BRS	brs.com
RLIN	rlin.stanford.edu
ORBIT	orbit.com

Uncover is a article delivery source, which covers about 15000 English language titles which can be accessed on the Internet. It has the most uptodate index. Access to the database is through journal title, title of the articles and author's name. It is managed by the famous publishing house Blackwell, Oxford. The web site is *http://www.blackwell.co.uk/libserv/technical/uncover.html*

15.5.5 Downloading the files

After searching the web and finding the information on the web, you may feel to download the files from the Internet to your computer. There is a lot of software available on the web either for free or as shareware. Shareware provides a part of the software for free for evaluating it, if you decide to keep using it you are expected to pay for it. There are also software programmes (for example, application software, games, anti-virus programmes and internet browsers) are updated from time to time and new versions are available on the web.

Different systems on the Internet provide different levels of access to their sites. Some sites allow their users free access to the full range of their services, while others restrict the users only browse to the menu options. For example, the publishers of many bibliographic and full-text databases charge for use of their databases.

Once you find the appropriate site you can download the files using the set procedures. Many files available for downloading are compressed and these can be recognised by their file name extensions (such as .Z, .gz, .zip, .arc, .lhz, etc.). You have to use programmes like PkUnZip, WinZip and Stuffit, which can uncompress many different compression formats.

A word of caution, be careful when downloading files from the Internet. Only take files from reputable sources. If you download a file from some unofficial archive, it may contain a virus or other software designed to damage your computer. Always better download the files to a floppy disk.

15.5.6 Information Services

With this background, let us now concentrate on the information services available through the Internet. If I say, Internet is the one stop-shop for an information services - it is not exaggeration. Governments, educational and research institutions, business firms, industries, banks, publishers, libraries, non-government organisations, and individuals are now offering a variety of information services through the Internet.

A number of activities you can do on the Internet. Leon and Leon (1998) have listed the following.

1) Send and receive e-mail
2) Visit any web site available on Internet
3) Read and post articles in newsgroups
4) Download files to your PC
5) Chat with other users online
6) Access online multimedia including radio and video broadcast
7) Play games with others online globally
8) Subscribe to electronic newsletters, e-mails, etc.
9) Join contests
10) Contribute articles and other materials
11) Do online shopping
12) Post your web sites (including personal web pages)
13) Create an e-mail ID and account for you
14) Use the e-mail remainder service
15) Find a person's details
16) Find an institution or organisation's details
17) Create your or your institution's/organisation's home page

15.6 INTERNET IN INDIA

In India, the union and state governments have launched their web sites on the Internet. The National Information Centre's web site is very informative and covers information about all the ministries, state governments, and union territories, major institutions, government policies, census data, current events and many more. Government corporations, national boards, federations, banks and stock exchanges are also offering very valuable information through the Internet. These include trade information, access to different business databases and online directories of corporate firms, etc.

15.6.1 Major Organisations on Internet

If you once surf the Internet, you will notice that it is full of educational and entertainment material. Almost all the foreign universities, research institutions, funding agencies like the World Bank, Food and Agriculture Organisation (FAO), NGOs share their valuable project information through the Internet. Most of the well-known publishers started offering their publications and journals in the form of electronic publications and electronic journals. Tata McGraw-Hill, Elsevier, Academic Press and Sage are now very popular electronic publishers on the Internet.

The newspapers and magazines all over the world have really gained due to the Internet. Therefore today, you will find *New York Times, The Tribune, The Hindu, The New Indian Express, Deccan Chronicle, Andhra Bhoomi* (Telugu) and *Siasat* (Urdu) on the Internet. You will also find popular magazines like *Time, Economist, India Today,* and *Frontline* on the web. The major news agencies also need to look at the Internet for reaching wider audience. Now, you can tune to BBC, CNN, All-India Radio (AIR) and Door Darshan on the Internet.

The Internet based information services are also being used for social and economic development. Today, we have concepts like *Government Online, Telemedicine, CyberDoctor, Wired Villages* and *Virtual Universities*, which are aimed at improving public administration, health, agricultural production and education in the remote places. A kind of commercial information service is now available online as *cyber malls*, wherein you can get whatever information you want. It may be about a tourist place or details of hotels or yellow pages or daily news. You can look for bright career opportunity over the Internet.

A few Indian web sites include

Reserve Bank of India	http://www.indiaworld.com/home/rbi/
The Hindu	http://www.indiaserver.com/news/thehindu/thehindu.html
Businessline	http://www.indiaserver.com/news/bline/bline.html
Hindu Online	http://www.webpaage.com/hindu/index.html
Indian Express	http://www.express.indiaworld.com/
Economic Times	http://www.economictimes.com/

15.6.2 Virtual Libraries

The Internet has tremendous impact on the libraries all over the world. As discussed earlier, you can now access library catalogues of all the major libraries in the world. Due to the Internet technologies the collection, organisation and retrieval of information in the libraries is going to change very rapidly. Now the libraries have been concentrating on information access and developing into virtual libraries that will provide access to any type of information available anywhere on the Internet. There is no physical presence of documents in their vicinity. Users can use the resources of these virtual libraries without physically visiting them. For example, the famous libraries available on web include -

British Library	http://www.portico.bl.uk/access/
Library of Congress	http://www.loc.gov/
Vatican Library	http://www.software.com/is/dig-lib/vatican.html

15.7 LET US SUM UP

Let us recapitulate briefly what has been discussed so far in this unit.

- The Internet is a network of thousands of computers scattered across the globe that allows free exchange of information. The Internet is a network of thousands of computers scattered across the globe that allows free exchange of information.
- The Internet was started in 1969 as defense project, namely, ARPANET, in the USA. It was developed by US Department of Defence's (DoD) Advanced Research Project Agency (ARPA).
- Internet uses physical connections usually telephone lines, direct wires, fibre optics, and satellite transmissions to link one computer network to another. It requires Internet protocols (TCP/IP), Subscription to Internet Service Provider (ISP), Browsers, etc to connect to the Internet.
- The Common Internet Services include - Information retrieval services (FTP and Gopher), Information Search Services (WAIS, Archie and Veronica), Communication Services (E-mail, Telnet, Usenet and IRC), and Multimedia Information Services (World Wide Web).
- For Searching the Internet, you need to know the web page address. There are search engines (such as altavista, hotbot, lycos, infoseek and yahoo), which help the users in finding the appropriate web sites for information.
- In India too, several governmental and non-governmental organisations as well as commercial firms have been creating their own web sites and using them.

15.8 REFERENCES AND RECOMMENDED BOOKS

CRUMLISH, Christian. *The ABCs of the Internet.* New Delhi: BPB Publications, 1996.

DICTIONARY of computer and Internet terms, 5th ed./ Douglas Downing, Michel A. Covington and Melody Muldin Covington. New York: Barrons, 1996.

Internet Library: *Case studies of library Internet management and use* /edited by Julie Still. Westport,CT: Mecklermedia, 1994.

KRAYNAK, Joe and Joe Habraken. *Internet 6 in 1.* New Delhi: Prentice-Hall of India, 1998.

LEON, A and M Leon. *Internet for everyone.* Chennai: Leon Tech World, 1998.

LEVINE, John R and Carol Baruodi. *The Internet for Dummies.* New Delhi: Pustak Mahal 1994.

LIBRARIANS on the Internet: Impact on reference service/edited by Robin Kinder. New York: The Hawthorne Press Inc., 1994.

SIWATCH, Ajit Singh. "Essence of Internet for libraries and information centres in electronic information era". *(IN ILA Seminar Papers: Libraries and information services in the electronic information era/* edited by J L Sardana, et al. Delhi: ILA, 1999).

15.9 ASSIGNMENTS

1) Listout at least ten web sites related to library and information centres.

2) Browse through at least two library web sites in India or any other country and give your impressions.

15.10 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) Describe how you establish a connection to the Internet.

2) Explain briefly the various Common Internet Services that are available on the Internet.

3) Write an essay on how you proceed with searching the Internet. Illustrate your answer with suitable exmples.

4) Discuss the role of different governmental and non-governmental organisations in making provision for Internet-based information services.

II SHORT NOTES

a) Virtual Libraries

b) Dial-up Networking

c) Gopher

UNIT - 16 : LIBRARY EXPERT SYSTEMS

Structure

16.0 AIMS AND OBJECTIVES

An Expert System is a computer programme that has the knowledge of experts in its knowledge-base. These computer programmes are capable of performing extremely specialised tasks and are helpful in problem-solving. This unit aims to introduce the concepts — Artificial Intelligence (AI) and Expert Systems and their application in information retrieval.

After studying this unit, you should be able to

- trace the origin and development of AI and expert systems
- describe the basic elements of AI and expert systems
- explain the computer programming systems in AI
- discuss the application of expert systems in LICs, with special reference to information retrieval (IR).
- list out various expert system search engines used in IR.

16.1 INTRODUCTION

We find that there is a certain level of expertise available among its practioners in a field. The experts and their expertise are used to advantageously in problem solving situations as well as to improve their products and services. Libraries and information centres are not an exception to this.

Performing certain complex tasks in organisations, such as LICs, requires the know-how of the experts and specialists. Unfortunately, there is usually a shortage of experts. If these experts leave their organisation, their knowledge or expertise is also taken away by them. Since we cannot clone such experts, we may find ways to use their expertise. The challenge is to capture and automate their knowledge to make it available to all those who need it. It is extremely difficult to succeed by using conventional methods of computer programming. However, the advances in information technology in recent decades, especially the use of artificial intelligence, made it possible to capture the knowledge of experts.

16.2 ARTIFICIAL INTELLIGENCE

According to *The New Shorter Oxford English Dictionary*, Artificial Intelligence (AI) is a field of study that deals with the capacity of a machine to stimulate or surpass intelligent human behaviour.

16.2.1 Origin and Development of AI

Artificial Intelligence is a field of computer science, which tries to explore the qualitative capabilities of the computers. By devising methods computers can used to perform tasks that require cognitive abilities. The potential of computers was foreseen by Alan Turing in 1950. He formulated a test (named after him), whereby a machine can be deemed to perform intelligently as humans.

The researchers of the same period tried to use computers to simulate the functioning of the neural nets of the brain. The field was named in 1956 by John McCarthy, who also equipped it with its *lingua franca*, a list-processing programming language called, *LISP*. In late 1950s, the research in AI concentrated on finding a general problem-solving techniques. The method of heuristic search or trial-and-error search was invented.

During the mid-1960s further progress in AI was achieved by creating knowledge-based systems. These systems contain interrelated information about a specific domain (a disease, for example) from which they can reason about that domain (render a diagnosis, for example). The reasoning process, called Inference, can be based either on a form of Heuristic Search or, preferably, on a causal model of the domain.

The knowledge-based approach has led to progress in the simulation studies on human cognitive faculties. By the late 1970s artificial intelligence had arrived as a discipline that was generating many practical results. In the early 1980s this work led to the establishment of the "fifth-generation research agenda" (originating in Japan), the goal of which was to organise computer systems that draw inferences rather than merely calculate. This 10-year research project began in 1982 with an estimated cost of $1.5 billion and aimed at developing the systems with a capacity of 1 billion logical inferences per second. This prompted United States and European countries to renew their interest in AI.

16.2.2 Basic Elements of AI

The fundamental approaches to problem-solving through AI are Heuristic Search (or discovery of a solution to a problem), Knowledge representation and Inference Engines.

The Heuristic search is applied to discover a solution to a problem through a search among alternative choices. The problem may be winning a chess game or proving a theorem in logic. It is done through the disciplined generation and evaluation of a number of promising possibilities. This trial-and-error process of moving toward a solution is employed. There is no algorithm that generates a solution directly.

Heuristic search paths are symbolically represented in the form of hierarchical structures, called a *Pyramid* or a *Tree*. The solution paths run from the initial stage (root node) along the branches of the tree and terminate on the leaves (terminal nodes) called the "goal state". Heuristic search limits the solution paths that are considered at any given point to the promising ones. Heuristics do not guarantee the optimally of a solution, however, they generally produce good enough solutions and render problems tractable. Important classes of problems — such as theorem proving, action planning for robots, or playing master-level chess — have been programmed using heuristic search.

Knowledge Representation is an important element of AI. The purpose of this is to organise required information in such a form that AI programmes can readily access it to perform cognitive functions. Generally, knowledge representation schemes are classified into *Declarative schemes* and *Procedural schemes*. First one refers to representation of facts and

assertions, while the later one refers to actions. The declarative schemes or object-oriented schemes include Relational schemes (Semantic network) and Logical schemes. Thus, a Knowledge base is created with semantic networks, in which nodes represent concepts and links between nodes represent relationships between the concepts. An example of this is the so called, IS-A link, which imposes a hierarchy on a network. Thus, if the system knows that a "dog" IS-A "mammal", the node "dog" inherits all of the properties of the "mammal" node, while only the speecialising properties of the "dog" need to be stored in the "dog" node.

Inference Engine is another component of AI systems. The structure of an Inference engine is related to the knowledge representation method. AI relies heavily on techniques of "inferencing" (generating new facts from existing data). Logical deduction is used as the rule of inference. A well-known proof method, called *Resolution*, proceeds by refutation; it combines the negated statement to be proved with a series of statements in the knowledge base; if a contradiction results, the original goal statement is proved to be true. Thus, inferences may be drawn and new knowledge produced. For example, AI uses this to explore the presence of an oil deposit at a given site; by a natural-language processing program to understand on utterance; or by a computer-vision system to interpret a scene.

16.2.3 Computer Programmes of AI

The computer programmes of AI are among the largest and most complex programmes ever developed and used. They consume a lot of time of the system designers and programmers. There are two basic general AI system development tools (programming languages), namely, LISP and PROLOG. These tools have been used largely to answer the programming requirements of AI and expert systems.

LISP is a practical list-processing programming language with recursive function capability for describing processes and problems. All programmes and data are written in the form of symbolic expressions and they are stored as list structures. The items represented in the lists are called atoms. In other words, LISP programmes are composed of atoms, groups of atoms (lists) and groups of lists (expressions).

PROLOG (PROgramming in LOgic) is a logic-oriented language developed by A.Colmeraner and P.Roussel at the University of Marseilles AI Laboratory. Further work on PROLOG was carried at the University of Edinburgh (Gt. Britain). It is basically a theorem-proving system. Originally developed for natural language processing. PROLOG programmes are composed of facts and rules.

16.2.4 Application of AI

In AI systems, computer programmes are developed to perform tasks, such as reasoning, adopting to new situations, and learning new skills. Through these programmes, people can work out how to use a tool they have never seen before, they can recognise faces, and they can learn new languages or how to diagnose diseases. The potential of AI applications is vast and it covers virtually every activity of human intelligence. AI is used mostly in natural language processing, text processing, speech synthesis, machine translation, computer-aided instruction, etc.

AI systems are applied in a number of problem solving situations. AI can offer a diagnosis within a medical speciality; analyse the structure of a chemical compound and suggest pathways for its synthesis; explore the presence of an oil deposit at a given site, predict weather conditions or manipulate manufacturing robots to perform a set of useful tasks.

16.3 EXPERT SYSTEMS

According to *The New Shorter Oxford English Dictionary,* an Expert System is "a computer program into which has been incorporated the knowledge of experts on a particular subject so that non-experts can use it for making decisions, evaluation or inferences." An expert system captures the knowledge and judgements of a specialist and makes their experience and expertise available to anyone with computer systems.

16.3.1 Need and Purpose of Expert Systems

Expert Systems are a branch of Artificial Intelligence. They are Knowledge-based systems and their database stores a description of decision-making skills of human experts in some domain of performance, such as medical image interpretation, taxation, configuration of computer system hardware, troubleshooting of malfunctioning equipment, etc. The purpose of designing expert systems is the desire to replicate the scarce, unstructured, poorly documented and empirical knowledge of a few specialists so that it can be readily used by others. Expert systems relieves the specialists from the routine tasks enabling them to devote more time to strategic growth of the organisation. Thus, expert systems can enhance decision-making power, ensure consistent and speedy decisions, improve job satisfaction and relieve pressure on experts, improve efficiency of workforce, reduce costs, save manpower, increase flexibility, etc.

16.3.2 Components of an Expert System

Expert systems have three major components:

1) a software interface through which the user formulates queries by which the expert system solicits further information from the user and by which it explains to the user the reasoning process employed to arrive at an answer;

2) a database (called the knowledge-base) consisting of axioms (facts) and rules for making inferences from these facts; and

3) a computer programme (dubbed the inference engine) that executes the inference-making process.

In Expert Systems the concepts and knowledge about the domain (knowledge-base) are stored as sets of facts and rules as follows: IF (condition) —> THEN (action)

The knowledge-base is an interconnected act of pattern-coded hypothesis, observations and rules developed by the Knowledge Engineer (a computer scientist, who specialises AI). He works hand in hand with a domain specialist and seeks to draw the required expertise

in order to create the knowledge base. He refines and structures the knowledge for use in expert systems. He also collaborates with the specialists for conducting user studies and user modeling.

The knowledge base is a linked structure of rules that the human expert applies, often intuitively, in problem solving. The process of acquiring such knowledge typically has three phases:

1) a functional analysis of the environment, users and tasks performed by the expert;

2) identification of concepts of the domain of expertise and their classification according to various relationships; and

3) an interview, either by human or automated techniques, of the expert(s) in action.

The results of the above steps are translated into so-called production rules (of the form "IF condition x exists, THEN action y follows") and stored in the knowledge base. Chains of production rules form the basis for the automated deductive capabilities of expert systems.

An Inference Engine, usually written in high-level programming language, acts as an Interpreter. Inference is performed through either forward chaining (moving from the conditions to the conclusion) or backward chaining (hypothesizing a conclusion and moving toward the appropriate rules and presented data). The ability to produce fruitful hypotheses (abductive reasoning), much like the human thought process known as a "leap of imagination".

Expert systems cannot handle unanticipated events, but they can evolve with usage. Current expert systems are incapable of inductive inference; that is to say, they cannot generalise. The generalisation process is controlled by meta-heuristic knowledge, i.e., knowledge about the formation of new knowledge). Because humans also learn much by analogy, researchers are investigating methods that will permit programmes to simulate human learning. Like humans, the best expert systems also learn from experiences.

Traditionally, library and information centres deal with classification, cataloguing, indexing, thesauri building, etc activities. With the introduction of information technology in LICs, the scenario of their activities and services have been changed completely. Now the library expert systems add still another phase of further development.

16.4 LIBRARY EXPERT SYSTEMS

In recent years, librarians too have started showing a lot of interest in developing expert systems to improve their services. Expert systems research is a field of AI that tries to design intelligent computer systems with built-in expertise of human specialists in a specific domain of knowledge and also use these concepts for problem solving.

16.4.1 Basic Elements of Library Expert Systems

The basic elements/components of AI (Section 16.2.2) and Expert Systems (Section 16.3.2) are applicable to library expert systems. Let re-examine these elements here in the context of library and information centres. Expert systems are computer programmes composed of a knowledge base (specific domain of knowledge) that contains the information supplied by a human expert (a specialist). An inference engine applies the appropriate information from the knowledge base to a specific problem. At present there are three common formats (namely, semantic nets, frames and production systems) for representing the information in a knowledge base. Let try to understand these concepts in the context of library and information science.

Semantic nets are the networks that denote objects and their relationships. A location with a library may form a semantic net. The objects within the location and their relations are described. Information about the objects in the semantic nets could be retrieved through execution of queries using an inference engine. (A partial representation of semantic nets is shown in figure-1)

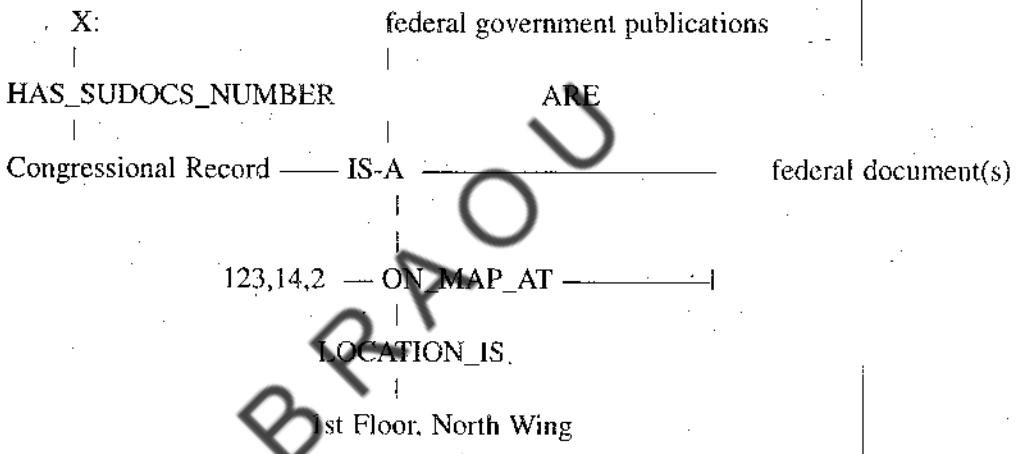


Figure-1: Semantic Nets
Source: Encyclopedia of Library and Information Science (Vol. 60.; p.203)

Frames are modular in structure. They lend themselves more readily to database encoding. For example, a specimen structure of a frame and its object information for the semantic net (figure-1) are as follows:

Generic LOCATION frame

Description:	NAME of requested location
Position:	DESCRIPTION of location.
Map coordinates:	X, Y on map display Z

The inference engine fills the slots (description, position, map coordinates, etc) in the frames in order to reach a solution to the problem. For example:

Documents LOCATION frame

Description:	Government publications
Position:	1st Floor, North Wing
Map coordinates:	123, 14 on map display 2

Production Systems are the most common forms of knowledge representation. They are composed of condition-action rules (IF — THEN pattern). The inference engine applies the appropriate rules from the knowledge base to solve a specific problem. Librarians unknowingly use these rules every day. For example, a user needs a journal article contributed by a certain author, but the location of the article is unknown. Then librarian would know to look in a periodical index. This may be expressed as an IF-THEN rule:

IF the item is located in a periodical
THEN consult a periodical index

16.4.2 Need for Expert Systems in LICs

Library experts generally use their common knowledge in developing, maintaining and expressing their expertise. In many situations problems are not that simple to fit them to mere IF-THEN expressions. Librarians have to use their innovation and creativity, besides their professional knowledge and experience to deal with the problems. They have to make new rules to suit the complex situations and develop useful expert systems using tools that are currently available.

Librarians have to deal with complex problems in administering their libraries and information centres on daily basis. They face problems with regard to administration (budgeting, staff, planning, etc), technical services (cataloguing, classification, collection development, etc) and user services. Expert systems could be designed and used to tackle these problems.

16.4.3 Use and Application of Library Expert Systems

Library expert systems could be helpful in assisting the librarians in making certain decisions in problem situations faced by them. They could develop systems composed of heuristics (rules-of-thumb) to make decisions. These rules could be weighted based on their effect. Let us examine some problem situations, where library expert systems are useful.

1) ***Budgeting:*** Expert systems are useful in the problem situations leading to IF there is a budget cut, THEN to determine — What items to cut and by how much? Shall staff be reduced? Should subscriptions to serials be cancelled? Which serials to be cut? Will there be any money to buy books or equipment? How much money be spared for maintenance?

2) ***Staff Management:*** Expert systems are helpful in recruiting, promoting, transferring as well as controlling the staff. For example, IF a staff member to be recruited/promoted,

THEN determine how much weightage to be given to qualifications, experience and references, and how to evaluate the quality and quantity of work produced?.

3) ***Planning:*** Expert systems have been using in designing library buildings, planning for remodeling, providing new facilities, etc. It also helps in determing the proper locations for circulation desk, photocopying equipment and OPAC terminals. Expert systems may be developed for planning of any task in a library, but one should keep in mind that it is expensive and time consuming.

4) ***Serials Management:*** Serials management in large libraries is a difficult task as their prices escalate from time to time, publication of new serials, issuing of serials in print and electronic formats, etc. Often librarians have to face serious problems with regard to budget cuts. For example, IF a serial is cancelled, THEN the faculty members will be upset. Librarians have to weigh the value of serial titles to the department, study its usage, calculate the cost involved, and other factors before making any changes in the subscriptions. Expert systems are extremely useful in assisting the librarians in making decisions.

5) ***Collection Development:*** Creating a collection development expert system for general academic libraries is a more complicated task. These libraries hold materials on varied subject areas for users of different educational pursuits in a number of schools/ departments. However, the attempts to build expert systems for processing of gifts were successful. Here, there are only two possible outcomes (yes or no) to the question of whether to keep a gift or not.

6). ***Cataloguing:*** Expert systems have been developed to create MARC records. Efforts have also been made to develop cataloguing expert systems based on the rules of AACR2, but it usefulness was limited because such systems had no means of interpreting the rules. In identifying authors, titles, series, etc., cataloguers use their personal experience in applying appropriate rules. Many experts feel that the rules in AACR2 are insufficient for developing a competent cataloguing expert systems. Therefore, much emphasis need to be given in developing the cataloguing rules. Research work is in progress for the development of systems for assigning subject headings. Some of the major projects of developing expert system cataloging tools include SKICAT (tools to analyse and explore large catalogue databases), CLARR (to assist in MARC field validation), Project DELICAT (to detect errors in library catalogues and drawing these to the attention of library staff), etc.

7) ***Classification:*** Classification of library materials requires broad knowledge in order to accurately assign call numbers. Usually librarians rely on title, table of contents, and indexes to ascertain the subject area. Expert systems find it extremely difficult to deal with new subject areas, non-representative titles, and synonymous terms. Therefore, research in this area is marginally successful.

8) *User Services:* Research is underway to develop expert systems for reference and information services. Expert systems would be useful to guide the users in locating materials and information. They would direct the users to specific sources of information to answer their queries and thus, they relieve the reference librarian from the routine chores of work. However, useful expert systems for reference are still in their infancy.

9) *Patent Search:* Another application of expert systems research is Patents Search. The Engineering Library of the University of Texas developed a system which guides the users through the process of a patent search. The Patent Search Tutorial is available through the Internet since 1997.

As an online intermediary, expert systems have constructed for searches and to select appropriate databases for non-experienced searchers. Let us discuss this in detail in the next section.

16.5 EXPERT SYSTEMS IN INFORMATION RETRIEVAL

Research has been carried out on the development of expert systems for information retrieval. Let us discuss about the expert systems developed in the field of online searching, automatic indexing, automatic summarising and subject analysis.

16.5.1 Online Searching and Expert Systems

Many expert systems have been created in recent times with an intention to provide assistance to searching databases or hosts. This will eliminate or release the human search intermediary from the more routine aspects. This has become necessary because users often encounter certain problems when formulating a subject search. Such common difficulties may be identified as:

1) Matching subject search terms to those found in the online or database system;

2) Increasing or decreasing their search results;

3) Understanding the organization of the database or online system, and

4) Recognizing the relevant documents, other than those retrieved, which have been scattered in the system due to interdisciplinary nature or other characteristics.

Studies on database searching reveal that about 60 percent of online catalogue searches are for subject information and difficulty in subject searching is the most important single factor in user dissatisfaction. The reason for this may be that most the existing database systems place the burden of search formulation, reformulation and evaluation on the user.

Traditionally several techniques are applied to assist the users in searching. The major ones are:

1) *Boolean Searching:* Using the words "and" "or" and "not" to narrow and expand search queries;

2) *Fuzzy Boolean or "Best Match" Searching:* All the documents containing search terms are displayed in ranked order, with the documents most closely matching the user's query shown first; and

3) *Automatic Search Sequencing* (Query Expansion): When an initial search does not produce adequate results, the system reformulates the search using its online thesaurus until desired results are achieved.

The above techniques help searchers to retrieve documents based on keywords of the user's query rather than on concepts. The retrieval systems do not provide instruction on the underlying principles of search formulation. This causes users' failure at searching and retrieval systems. Therefore, expert systems have tremendous potential to aid users with these more complicated issues.

The need for development of expert systems to assist the users have been realised by the database designers and information scientists. Much attention has been paid in recent times to develop expert systems as intelligent front ends and back ends for online databases. With the creation of expert systems the online databases would become more user-friendly. These systems assist users with the intellectual processes (i.e. conceptualisation of the search topic, structuring of the search logic, and interpreting and refining results) of online searching. The crux of the problem lies in guiding the users through their search strategy. The expert system engages the users in the search process by presenting alternatives to the search words the user has chosen, surveying the user's satisfaction with the output and then using that information to guide him in narrowing or broadening his search till he gets desired information from the search. This resulted in the creation of expert intermediary systems.

An expert intermediary system or expert retrieval assistance system is one that embodies within it the knowledge and skills of a librarian or "search intermediary" for carrying out online searches in bibliographic databases. An expert intermediary system differs from the other kinds of expert systems in two aspects:

1) The expert intermediary system is concerned with indirect access to information. Its expertise is centered on the techniques for retrieving references to documents rather than actually deducing and providing facts.

2) The domain or subject coverage of the retrieval system is usually wider than for a typical expert system.

Many expert intermediary systems have been developed for information retrieval purposes. They deal with abstracts or articles of full-text journals.

Examples of Expert Systems used in Online IR

Some examples of existing or developing applications of AI and expert systems to online searching include:

1) CANSEARCH

CANSEARCH is a rule-based expert system designed to aid doctors in searching the MEDLINE database for information on cancer therapy. The system's success is attributed to its menu-driven touch screens, its limited domain and its expert user population.

2) ANSWERMAN

Answerman is a ready-reference system developed by the National Agriculture Library (USA) that uses forward chaining to provide answers to user queries from a knowledge-base consisting of facts from agriculture reference books.

3) ORA

ORA is a system that uses forward chaining to pose a set of questions to the user and recommend an appropriate reference source. This system was tested at the University of Waterloo.

4) PLEXUS

PLEXUS is an expert system developed by the University of London, designed to direct users toward literature on gardening. Queries are entered in natural language and revised automatically and through user-system dialog. The system collects data about the user's searching ability and gardening experience to incorporate user modeling into its interactions.

5) PR

PR is a system designed to provide a user with a variety of means to access information from a database on computer science. The user provides a Boolean search, a relevant document or a natural language inquiry. In turn, the computer presents a list of relevant search terms for the user to choose from. If desired results are not achieved, the system is able to reformulate its technique to include related terms.

6) SCISOR

SCISOR is a system designed to match natural-language inquiries about corporate acquisitions and mergers and match them with conceptual representations of news stories in its database. This system is attached to the Dow Jones online financial service.

16.5.2 Expert System Search Engines

There are several expert system search engines developed to assist the users of online searching and information retrieval from the databases. Let us study the major search engines

that are in vogue. (Information about the AI/Expert Systems and the search engines has been retrieved from the Internet sites using the search engine: *http//www.mamma.com)*

1) Excite

The Excite search engine is powered by Intelligent Concept Extraction (ICE), a technology that uses advanced algorithms to retrieve documents and score them based on relevancy of their concepts. Excite's classification scheme is driven by statistical analysis of the documents themselves and the assumption that words that frequently appear together are related. When a new document appears on the Web, ICE analyzes the relationship between the site's text words and learns news associations. ICE uses the associations to retrieve documents containing concepts relevant to user queries, whether or not that document contains the search keywords.

Excite reports that its method is as effective in terms of recall and precision as Latent Semantic Indexing, but is faster and more efficient. The amount of time necessary for Excite to perform a search will remain constant, even as the web grows exponentially.

Additional features of the ICE system include:

* Query By Example, which allows users to access documents that are similar to a document they find particularly relevant;

* Automatic Subject Grouping, which allows users to clarify their search topic if their keyword has multiple meanings;

Automatic Abstracting, the system's ability to select sentences that are relevant to a document's concept and display them to aid the user's process of evaluation.

2) HotBot

Inktomi, the search engine behind HotBot, is one of the first real-world applications to exploit multi-machine parallel computing. Other systems are constrained by the storage of their database on one large computer. If an increasing user base and body of information necessitates the addition of another computer, the information on the first computer must be duplicated on the second. By employing a database system whose information is designed to be stored in several linked computers, Inktomi avoids this problem. The result is a fast, low-cost system with outstanding growth potential.

Inktomi is able to produce detailed reports for its advertisers by tracking the sites visited most often by its users. The information is used for data mining (INSERT) and decision support analysis.

3) ConText

Oracle's ConText uses a combination of hierarchical indexing, natural language processing and machine learning to automatically classify web documents. ConText has devoted more than 100 person/years to building a knowledge classification scheme, and supplemented the resulting information with automatic statistical and document analysis techniques.

16.5.3 Automatic Summarizing

Another application of expert systems research is Automatic Summarising. Automatic summarising is the process by which a computer creates a condensed version of a text, The new version, thus created, should well represent the original in meaning, scope and content.

The need for automatic summarising has been felt with the explosion of literature. The libraries are often overwhelmed by the volume of material they acquired. They must catalogue and make available to the public in a short tir .e. A computer programme that can write coherent and meaningful summaries of textual material is a solution to the problem.

Much of Automatic Summarising relies on Natural Language Processing. Natural Language Processing is what a computer does when it makes syntactical and semantic sense of non-mathematical language. A computer that can do this must understand abstract linguistic formulae as well as general meaningful contexts. Artificial Intelligence programmes use different techniques to analyse the natural language. The major techniques used in analysis and generation of responses are briefly discussed below:

One technique involves instructing the computer to look for keywords and manipulate them according to pre-programmed pattern rules.

Another technique instructs the computer to analyse the way the words in each sentence are related to one another. Then, each word is assigned a meaning according to the semantic knowledge previously applied to the computer.

A third technique relies mostly on knowledge. The computer examines the text and determines what category of action each sentence (or idea) falls into. Then, the computer makes sense of the text under broader domain knowledge it has already been supplied.

The result of all these operations is that the computer generates a response in natural language that makes sense grammatically and semantically to the user.

Automatic Summarising systems draw knowledge from a variety of disciplines, besides computer science. Computer science, of course, provides the skills needed to use the computer to the best of its technical ability. A knowledge of human cognition helps us understand how we summarize, thereby suggesting structures we may be able to apply to a computer. Research in the field of linguistics, especially grammatical structure and domain knowledge, has been part of the foundation of Automatic Summarising systems. Investigation into the way we psychologically understand text has addressed the complex problem of context and computer comprehension.

The success of the machine generated summary depends much on different factors, such as the characteristics of the material and the purpose of the summary.

The essential characteristics of the material to be submitted to the computer for automatic summarising are:

FORMAT (size and structure of the text)

TOPIC (the general topic and the degree of specialization)

NUMBER (whether the input is a single item or a set of items)

The purpose of the summary:

CONTEXT (the situation in which the summary will be used)

AUDIENCE (the people who will be using the summary)

FUNCTION (what the summary will be used for)

The result of work done by the automatic processing system:

CLASSIFICATION (the kind of material it is)

PRESENTATION (the way the summary is formatted)

LINGUISTIC STRUCTURE (the grammar of the summary)

In order for Automatic Summarising systems to be more successful, they must be user-friendly. Research on adapting interactive programs to users is in progress.

16.5.4 Automatic Indexing

In subject analysis and indexing certain words or phrases suitably represent its information content of a document are selected and these sets of words are transformed into a standard terminology or assigned code, if necessary. The terms or codes thus selected are treated as access points for the information search. In manual indexing human analysists need to understand the topic or content of information in the documents.

The automated systems rely on the extraction and manipulation of keywords. They employ two types of rules in subject indexing :

1) the human analylist compiles a list of keywords of potential interest; this list is compared by the computer with each word in the text of the document; if a keyword appears, the fact is recorded, and these selected words comprise the index entry for the document; and

2) the human analysist compiles a list of words that are NOT to be selected for indexing (such as stop words); automated applications then count the number of times specific words appear in a document, and the most frequently-appearing words are selected for indexing.

The automatic indexing systems which were tested in the 1960s were partially successful when the queries were straightforward and domain-specific, and the collections had to be small.

The systems of the 1970s focussed mainly on the inquirer's approach rather than the power to manipulate documents. Designers of automatic indexing systems need to understand how questions were asked and what assumptions lay behind those questions. An other inherent problem with the indexing systems is that of the context dependent meaning of certain words. Expert systems have been developed taking these aspects into consideration.

One promising field of AI and expert systems research with implications for online searching is *Latent Semantic Indexing*. Latent Semantic Indexing is an approach used in natural language processing. The expert systems create a geometric representation of words within a document, reduces these symbols to a matrix and then analyses them mathematically to find correlations between the words. These correlations are used to find the documents that are related in concept to documents requested by a searcher, whether or not they contain the original keywords.

16.6 LET US SUM UP

Let us recapitulate briefly what has been discussed so far in this unit.

* Artificial Intelligence (AI) is a field of study that deals with the capacity of a machine to stimulate or surpass intelligent human behaviour.

* The fundamental approaches to problem-solving through AI are Heuristic Search (or discovery of a solution to a problem), Knowledge representation and Inference Engines.

* Expert Systems are a branch of Artificial Intelligence. They are Knowledge-based systems and their database stores a description of decision-making skills of human experts in some domain of knowledge. Expert systems are applied in the fields, such as medical image interpretation, taxation, configuration of computer systems, troubleshooting of malfunctioning equipment, etc.

* Expert systems can be applied to various problems faced by librarians with regard to administration (budgeting, staff, planning, etc), technical services (cataloguing, classification, collection development, etc) and user services.

* An expert intermediary system or expert retrieval assistance system is one that embodies within it the knowledge and skills of a librarian or "search intermediary" for carrying out online searches in bibliographic databases.

16.7 REFERENCES AND FURTHER READING

ELLIS, David. *New horizons in information retrieval. London*: Library Association, 1990. p.71-96

HOLTHOFF, Timothy. "Library expert systems". IN *Encyclopedia of library and information science/*edited by Allen Kent et al. New York: Marcel Dekker, 199_. Vol.60; p.333-5.

NEW Encyclopedia Britannica. 15th ed. New York: Encyclopedia Britannica Inc., 1993. (Macropedia: Vol.21; p.627-637)

*NEW Shorter oxford English dictionary: on historical principles/*edited by Lesley Brown. Oxford: Clarendon Press, 1993.

WORLD Book Encyclopedia. London: World Book Inc., 1992. Vol.I; p.637.

16.8 MODEL EXAMINATION QUESTIONS

I ESSAY QUESTIONS

1) Define Artificial Intelligence and describe its origin and development.

2) Explain the basic elements and computer programming systems of AI and expert systems.

3) What do you understand by Library Expert Systems ? Explain its application to various operations in LICs.

4) Trace the role of expert systems in online information searching and retrieval.

II SHORT NOTES

a) Search Engines

b) Expert Intermediary Systems

c) Latent Sema..tic Indexing