CONTENTS

SYLLABUS

INFORMATION PROCESSING AND RETRIEVAL

M.Lib-04

Unit – I:

DOCUMENT DESCRIPTION : PRINT & NON-PRINT MATERIALS: Bibliographic Description - An Overview - Standards for Bibliographic Record Format - Bibliographic Description of Non-Print Materials.

Unit – II:

SUBJECT ANALYSIS AND INDEXING SYSTEMS: Classification Systems : General & Special - Universal Decimal Classification - Subject Indexing - Automatic Indexing and Machine Translation - Thesaurus : Its Structure, Functions and Construction.

Unit – III:

INFORMATION STORAGE & RETRIEVAL (ISAR) SYSTEMS: Information Storage and Retrieval Systems -An Overview - File Organisation in ISAR Systems - Evaluation of ISAR Systems – Methodology - Evaluation of ISAR Systems - Experiments & Case Studies.

Unit – IV:

INFORMATION ACCESS AND RETRIEVAL METHODS: Information Access: Online and CD-ROM Databases - Database Searching: Search Strategies - Information Retrieval Protocol: Data Mining and Data Warehousing - Library Expert Systems.

References

1. Library Information Procedure, Edited by Shyama Balakrishnan and P.K. Paliwal, Anmol, 2001
2. Library Information and Society, Kusum Verma, Vista International, 2005
3. Management of Library Information Services, Edited by Shyama Balakrishnan and P.K. Paliwal, Anmol
4. Managing Library Information Systems, Edited by S.R. Das, Arise Pub, 2008
5. Information System, Edited by S.P. Singh, Omega Publication, 2008.

NOTES

UNIT I DOCUMENT DESCRIPTION —PRINT & NON-PRINT MATERIALS

★ STRUCTURE ★

LEARNING OBJECTIVES

After going through this unit, you will be able to:

- describe the future of print media
- know more about non-print media
- explain standards for bibliographic record format.

1.1 INTRODUCTION

Men have been communicating with speech for about 100,000 years. This form of communication, *i.e.*, the oral form, reigned unrivalled for thousands and thousands of years. Gradually, it dawned on man that message can be left on some surface using drawings or symbols. The great cave paintings of Altamira in Spain and Lascaux in France daubed on the cave walls some 20,000 years ago seem to convey some distinct message [*Odhams*]. Millennia passed by before drawings and paintings took the form of early pictorial writing. In some parts of the world pictorial writing gave birth to scripts. The oldest known writing we are aware of is that of Mesopotamia inscribed as early as 3,000 BC by the Sumerians. That only points to the fact that the practice of writing emerged only about 5,000 years ago. Printing that opened the floodgate for the production of books and revolutionised

the spread of education came into being much later. Evidences suggest that the Chinese invented the method of block printing by 8th century AD. Their method remained more or less confined to China and did not spread all over the world and the production of books did not attain the necessary momentum.

In 1450s (according to some source 1454 AD) Johannes Gutenberg of Germany invented the method of printing using movable types. The impact of this method of printing was of unimaginable dimension. The method spread like wild fire and by the end of the 15th century some 9,000,000 books were in circulation in Europe, a scene the world has neither seen before nor after [*Odhams: p55*]. At one stroke, the world saw the birth and development of the printing industry, printing machinery industry, publishing and book trade industry, printing ink industry, sudden rise in pulp and paper industry, and so on. It also created in people a tremendous urge for reading leading to the growth of literacy, educational institutions like schools, colleges and universities. Moreover, it gave rise to professionals like printers, composers, proofreaders, publishers, book traders, book binders, and so on.

The print media reigned supreme and unrivalled for about 500 years when at the 2nd half of the 20th century it faced a formidable challenge from non-print media.

Now, a big question has cropped up before the world whether the print media will be able to withstand the threat from the non-print media and continue as usual in future, or it will yield to the pressure and gradually vanish from the scene.

Print Media

Printing involves a minimum of four different items:

(*i*) Manuscript, *i.e.*, the piece of writing to be printed;

(*ii*) Composition of the matter either by hand or by machine;

(*iii*) The physical medium, say, paper on which the matter is to be printed; and

(*iv*) The ink with which the matter to be printed. For illustrations, blocks, etc., are also required. Products of printing are many and varied. For example, books, periodicals, newspapers, etc., are all products of printing and all of them represent one medium or the other.

All these products taken together from the print media. Hence, in this unit we are using the term print media instead of print medium.

1.2 BIBLIOGRAPHIC DESCRIPTION: AN OVERVIEW

Organization of bibliographic data elements leads to the creation of bibliographic records. Bibliographic record has been defined as the

NOTES

sum of all the areas and elements, which may be used to describe, identify or retrieve any physical item of information content. Bibliographic description is the assemblage of data elements sufficient to identify a bibliographic item and to distinguish it from others. In manual systems (*e.g.*, card catalogue), a collection of bibliographic data elements are grouped under the main access points or headings as per the cataloguing code in use. Such record of an item in a catalogue is called an 'Entry'. Entries are usually identified by the kind of access they provide *e.g.*, 'author entry' or 'subject entry'. The distinction between bibliographic record and entry is most visible in computerised environment where the master bibliographic record is stored in the machine and computer programmes generate entries from it. Dempsey [1989] identified three groups of bibliographic dataset – bibliographic description and control data (data describing, identifying and providing controlled access to items), subject data and content description. The first two groups of data generally appear in library catalogues and bibliographic databases. They include:

- data naming an item (*e.g.*, title, alternative title);
- data naming persons or bodies connected with the creation of an item (*e.g.*, author, artist, cartographic agency);
- data describing hierarchical, lateral or lineal relationships between items (*e.g.*, component part, host item, numbering in series, companion item, name of earlier edition or version);
- data indicating intellectual content (*e.g.*, subject heading, abstract);
- data naming persons or bodies connected with the production of an item as a physical object (*e.g.*, publisher, designer);
- data indicating form or nature of item (*e.g.*, bibliography, documentary, novel);
- data indicating mode of expression or communication (*e.g.*, verbal, pictorial);
- data describing the physical appearance, characteristics and constituents of an item (*e.g.*, map, film, dimensions, number of volumes or parts, technical information needed for use); and
- data assigned by a bibliographic or other agency for purpose of identification and control (*e.g.*, ISBN).

The above list shows that bibliographic description deals with two categories of data—data providing access and data describing items. The level and extent of bibliographic description depends on the application and purpose of bibliographic records. The major application domains are the production or creation of:

- authoritative national records and national bibliography;
- bibliographic records for international exchange;
- bibliographic records for cooperative systems;

- records for use in individual libraries;
- records for abstracting and indexing services;
- records for subject bibliographies/author's bibliographies;
- records for use in online information retrieval systems (including WWW); and
- records in book trade.

Bibliographic record may be viewed as a package of data, the content of which varies according to the different needs and purposes for which it is intended. The selection and inclusion of data elements for the bibliographic record must be based on user needs. The aggregate of data in a bibliographic record are broadly divided into following groups:

- Descriptive data elements (as defined in the ISBDs);
- Data elements used in headings for persons, corporate bodies, titles and subjects. They function as filing devices or index entries;
- Data elements used to organise a file or file of records (such as classification numbers, abstracts, summaries or annotations); and
- Data specific to the copies of the library collections (such as accession numbers and call numbers).

Bibliographic record should be constructed according to the agreed rules and standards. There are many widely used standards for constructing bibliographic records (*e.g.*, AACR2 for national bibliographies or library catalogues) but the most striking contribution has been made by IFLA, with its programme of ISBDs. ISBD(G) [Genereal International Standard Bibliographic Description] is intended to provide the generalised framework for descriptive information required in a range of different bibliographic activities. The bibliographic data elements which are required for this purpose are set out in eight areas: Title and statement of responsibility – Edition – Material specific data – Publication, distribution data – Physical description – Series – Notes – Standard number and terms of availability. Each of these areas is further divided into discrete elements. The elements are cited in given order and separated by the punctuation prescribed. The complete set of ISBD data is sufficient to ensure identification of bibliographic item and many cataloguing codes (including $AARC_2$) have adopted ISBD(G) as a basis for their own rules for description. The family of ISBDs (includes standard for cartographic materials, non-book materials, printed music, antiquarian books, monographs, serials and other continuing materials and electronic resources) is utilised for the purpose of bibliographical description but the choice and form of access points are based on the Paris Principles (the statement of principles adopted at the International Conference on

NOTES

Cataloguing Principles held in Paris in 1961) adopted in national cataloguing rules. Bibliographic formats (such as MARC family, CCF, UNIMARC, etc.) have also applied ISBDs as base format. But cataloguing codes and bibliographic formats cannot represent all the characteristics of different digital information resources. As a result various general and domain specific metadata schemas (such as Dublin Core, FGDC, ONIX, GILS, etc.) have been developed for description of electronic resources.

Scope and Objectives of Bibliographic Description

Bibliographic description is a tool for bibliographic control. The design of such a tool should begin with a statement of objectives. The objectives for library catalogues were first formulated by C.A. Cutter in 1876 and remained unchallenged for more than 75 years. The first suggestion of revision came in 1953 from Symour Lubetzky that calls for collocating various physical manifestations of a work, such as different editions and translations of it. In the last 50 years (since Lubetzky's reformulation of the objectives of catalogue) not only have catalogues been automated, but also an unprecedented amount of co-operative catalouging has led to the emergence of international standards, global catalogues, and linked systems along with vast array of digital resources available in Internet. Under such circumstances, Elaine Svenonius in her seminal book [2000] established that bibliographic systems are based on five objectives: finding, collocating, choice, acquisition and navigation. Actually this set of objectives is an expanded version of Functional Requirements for Bibliographic Records (FRBR) objectives. The Joint Steering Committee for Revision of AACR has adopted the propositions of Svenonius as a set of objectives for full-featured bibliographic system. These objectives may better be explained against backdrop of the scopes of bibliographic description in a rapidly changing cataloguing environment.

Scope

Bibliographic description should facilitate the scope of relating the data that are encoded in bibliographic records to the needs of the user of those records. It should ensure a basic level of functionality for records created by bibliographic agencies and libraries. However, the scope of bibliographic description may be discussed in the context of coverage, user, application and generic tasks associated with it.

(*a*) **Coverage.** Bibliographic description should be comprehensive in terms of variety of materials (textual, musical, cartographic, audio-visual, graphic, three dimensional materials, etc.). It should cover the full range of physical media of bibliographic records (paper, film, magnetic, optical storage media, etc.) and all formats (books, sheets, discs, cassettes, cartridge, etc.) along with all modes of recording (analogue, acoustic, digital, etc.).

NOTES

(*b*) **Users.** Bibliographic description should ensure the use of bibliographic records by a wide range of users. The user groups to be supported include readers, students, researchers, library staff, publishers, distribution and subscription agents, retailers, information brokers, administrators of intellectual property rights, etc.

(*c*) **Applications.** Bibliographic description vis-à-vis bibliographic records should support a wide variety of applications both within and outside a library setting in which the data in bibliographic records are used. The applications of bibliographic record should include (but not limited to) the following activities: collections development, acquisitions, cataloguing, production of finding aids and bibliographies, inventory management, preservation, circulation, inter library loan, reference, and information retrieval.

(*d*) **Tasks.** Bibliographic description should fulfill the functional requirements for bibliographic records. It must allow users to perform all the generic tasks at the time of searching and making use of bibliographic records. These tasks include:

- use of bibliographic data to find materials against the user's stated search criteria;
- use of bibliographic data to identify an entity;
- use of bibliographic data to select an entity that is appropriate to the user's needs; and
- use of bibliographic data to acquire or obtain access to the entity described.

Objectives

The objectives of a full-featured bibliographic system may be viewed in the historic perspectives of Cutter's objectives, Lubetzky's objectives (as reflected in the Paris Principles) and the modern groundbreaking works like IFLA's Functional Requirements for Bibliographic Records (FRBR) objectives and objectives propounded by Elaine Svenonious. The principles of bibliographic description should facilitate achievement of following objectives of a full-featured bibliographic system. The objectives are:

To enable a user

(*a*) to locate resources in a file or database as the result of a search using attributes or relationships of the resources:

(*i*) to find a singular resource;

(*ii*) to locate sets of resources representing;

- all resources belonging to the same work;
- all resources belonging to the same expression;

NOTES

- all resources belonging to the same manifestation;
- all resources by a given creator of intellectual or artistic content;
- all resources on a given subject;
- all resources defined by 'other' criteria (such as language, country of publication, publication date, physical format, etc.);

(*iii*) to explore bibliographic relationships (*e.g.*, to find resources which are bibliographically related to a given resource);

(*b*) to identify a resource or agent (*i.e.*, to confirm that the entity described in a record corresponds to the entity sought or to distinguish between two or more entities with similar characteristics);

(*c*) to select a manifestation or specific item that is appropriate to the user's needs (*i.e.*, to choose a resource that meets the user's requirements with respect to content, physical format, and so on or to reject a resource as being inappropriate to the user's needs);

(*d*) to acquire or obtain access to an item described (that is to acquire an item through purchase, loan, and so on or to access an item electronically through an online connection to a remote source);

(*e*) to navigate a bibliographic database (that is through the logical arrangement of bibliographic information and presentation of clear ways to move about, including presentation of relationships among attributes).

The above-mentioned objectives are applicable to both description and access of bibliographic entities. Cataloguers perform the following tasks to meet the above user driven objectives and also to maintain the catalogues or bibliographic databases:

- Transcribe
- Describe
- Make identifiable
- Link
- Manage
- Convey rights management information

Modern integrated library systems also enable the online catalogue to be more than a by product of the online work of cataloguers. It now extends support to integrate library operations to take advantage of global network access to other catalogues and bibliographic resources, to reach other bibliographic databases, to access virtual copies of resources in addition to providing expanded search, retrieval and display capabilities.

Principles of Bibliographic Description

Principles of bibliographic description are guidelines for the design of a set of rules to manage bibliographic data. In ordinary usage, these two concepts (principle and rule) are used many a time interchangeably. In fact, principles guide the creation of rules and thereby distinct from bibliographic objectives and bibliographic rules. The term 'principle' is used to refer to a proposition or other formulation, usually generalised and with one or more of the following attributes [*Bhattacharya, 1979*]:

- It may be a statement of fact;
- It may be accepted as true or helpful;
- It may form the basis for deriving another proposition with one or more of the attributes of the basis proposition;
- It may provide a basis for reasoning or evaluation; and
- It may guide the formulation of a proposition prescribing a procedure for fulfilling a particular purpose.

Principles

The Joint Steering Committee (JSC) for revision of AACR [Huthwaite, 2001] developed two broad groups of principles related to bibliographic description. These principles are profoundly based on the works of S. R. Ranganathan [1955], G.W. Leibniz [1951] and E. Svenonius [2000]. These principles of bibliographic description are designed to support an expanded role of library catalogues and bibliographic databases and therefore should be integrated with a set of principles related to bibliographic relationships. The principles of bibliographic relationships, as proposed by Velluci [1997], provide a logical direction for the treatment of bibliographic relationships in the universe of bibliographic entities and attributes.

- General Design Principles

  (*a*) Principle of Sufficient Reason (Based on Leibniz and Ranganathan's Law of Impartiality)

  Each design decision must be defensible and not arbitrary.

- Principle of Parsimony (Based on Ranganathan, S.R.)

  (b) When there are alternative ways to achieve a design goal, prefer the way that best furthers overall economy.

  Principles of Bibliographic Description and Access

- Principle of User Convenience

  (*c*) Decisions taken in the making of descriptions and controlled forms of names for access should be made with the user in mind.

NOTES

- Principle of Common Usage

  (*d*) Normalised vocabulary used in descriptions and access should accord with that of the majority of users.

- Principle of Representation

  (*e*) Descriptions and controlled forms of names for access should be based on the way an information entity describes itself.

- Principle of Accuracy

  (*f*) Descriptions and controlled forms of names for access should faithfully portray the entity described.

- Principle of Sufficiency and Necessity

  (g) Descriptions and controlled forms of names for access should include only those elements that are bibliographically significant.

- Principle of Standardisation

  (*h*) Descriptions and controlled forms of access should be standardized, to the extent and level possible.

- Principle of Integration

  (*i*) Descriptions for all types of materials should be based on a common set of rules, to the extent possible.

  Principles of Bibliographic Relationships

- Principle of Relationship Identification

  (*j*) The bibliographic record should identify all important bibliographic relationships that exist between the entity being cataloged and other entities. These relationships include both independent and dependent relationships. Identification should be bi-directional.

- Principle of Enabling Linkage

  (*k*) The data elements of the bibliographic record should enable related bibliographic records to be linked, and should permit the bibliographic record to be linked to related documents. To this end, the bibliographic record should provide enough information to identify the relationship and create a linkage. Linkages between bibliographic records should be bi-directional.

- The Principle of Multi-level Description

  (*l*) The cataloguing code should provide for the independent description of an entity at several levels, including the abstract work, the physical item, and the specific copy. These hierarchically related descriptions should be linked.

- Principle of Consistency

  (*m*) The identification and linkage of like bibliographic relationships should be treated in a consistent manner, regardless of physical format. This includes the consistent application and use of uniform titles.

NOTES

Bibliographic Formats

Bibliographic formats have been created for two purposes – to facilitate search and retrieval of bibliographic records, locally and in network environment and to exchange bibliographic information among library and information centres. A bibliographic format that acts as a means of exchanging data has three basic components:

Physical structure: It may be considered as a container or carrier of bibliographic data on a computer storage medium. ISO 2709 – an international standard of bibliographic record structure is accepted by the information community for the exchange of bibliographic data on magnetic tape and other storage media.

Content designators: These are codes to identify different data elements in the record and represented in bibliographic formats by tags, indicators and sub-field codes. There are many standard content designator schemes, which can be used to create and exchange bibliographic records such as MARC family (USMARC, CANMARC, UKMARC, UNIMARC, INDIMARC, MARC21 – a combination of USMARC and CANMARC, etc.), CCF (Common Communication Format), MIBIS (Microcomputer-based Bibliographic Information System) and others.

Content: The form and content of data elements should be based on some rules and codes. Here, library community is benefited by catalogue codes and ISBDs and community of abstracting and indexing services is benefited by UNISIST Reference manual.

Bibliographic description in machine-readable form requires a standard format to manage different types of bibliographic items, to cover variety of library and information services over a wide range of institutions and to support different computer configuration and programming languages.

Electronic Resource Description

Electronic resources in general and Internet resources in particular have some specific characteristics that call-for some special provisions in their description. ISBD(ER) provides a long list of data elements for describing electronic resources and AARC2 guides the cataloguing of such resources. In 1999, OCLC produced a manual for cataloguing Internet resources on the basis of AARC2 and ISBD(ER). Some bibliographic formats like UNIMARC and MARC21 developed a special field (Field 856) for managing electronic resources.

It has already been mentioned that bibliographic formats and catalogue codes are not adequate enough for representing all the unique characteristics of digital resources. As a result, various metadata standards have been developed as resource description schemas for digital information bearing objects over the past few years. Metadata is structured information that describes, explains, locates or otherwise

NOTES

makes it easier to retrieve, use or manage digital information resources. Metadata schemas are sets of metadata elements designed for a particular purpose, for example, to describe a particular type of information resource. In addition to resource discovery, metadata schemas can help to organise electronic resource, facilitate interoperability and resource integration, support digital identification and ensure archiving and preservation. Many different metadata schemas are being used in library environment. A few of the most common ones are mentioned here.

It was developed in 1995 to be simple and concise scheme, and to describe web based documents. The original objective of the Dublin Core was to define a set of elements that could be used by authors to describe their own resources. Dublin Core is a set of 15 main elements that fall into three groups–Contents, Intellectual Property and Instantiation. Simple DCMS applies only main 15 elements without any qualifier. Qualified Dublin core uses additional qualifiers to increase specificity or precision of the metadata.

(*a*) Global Information Locator Service (GILS) [http://www.usgs.gov/gils/index.html]

GILS grew out of U.S. government requirement for public access to government information, – both digital and non-digital. The National Archives and Records Administration, U.S. has defined the core elements of GILS. GILS specifies a profile of the Z39.50 protocol for search and retrieval. GILS records are intended to describe aggregates such as catalogues, publishing services and databases.

(*b*) Text Encoding Initiative (TEI) [http://www.tei-c.org]

It is a scheme for marking up electronic text. It also specifies a header portion to accommodate metadata about the object to be described. TEI headers can be used to record bibliographic information of both electronic and non-electronic sources. The TEI header can be mapped to and from MARC.

Evolution of Bibliographic Description

Standards for bibliographic description have existed in one form or another for well over a century and have been subjected to change either by the evolution of cataloguing theory or by force of practice throughout the entire period of their existence. The different cataloguing codes starting from Panizzi's Rules for Descriptive Cataloguing are examples of this. The application of ICT and the increasing popularity of shared cataloguing have influenced the development of standards related to bibliographic description over the last 35 years. The standards, rules and principles of bibliographic description are also applicable to online and web environment for discovery, indexing and identification

NOTES

of digital resources by semantic means. Metadata schemes for describing electronic resources (such as Dublin core, FGDC, GILS, etc.), like traditional bibliographic standards and formats, are also designed for information storage and transfer. It means that a bibliographic record now performs different functions with respect to various media, various applications, and various user needs. This situation calls for a framework that should identify and clearly define bibliographic entities, attributes, relationships and tasks performed by users of bibliographic records. Some conceptual models (such as IFLA's FRBR model, UKOLN's Analytical Model of Collection Description and XOBIS model of Stanford University) have been developed in recent years to encompass a broad-spectrum application of bibliographic description and use of bibliographic records. The growth and development of bibliographic description standards are discussed below:

Pre-ISBD Development

Four landmarks–Panizzi's code, Cutter's contribution, Ranganathan's scientific principles and Paris principles, characterise this period of evolution which ranges from 1841 to 1969. Anthony Panizzi, then librarian of the British Museum, is regarded as the first person who codified rules for cataloguing by preparing a formal code of rules for cataloguing in 1841. In 1876, Charles Ami Cutter provided a direction in cataloguing through a set of well-defined objectives and Ranganathan first formulated a set of normative principles to provide scientific basis for bibliographic description. Paris Principles (1961) are first international agreement on general rules of cataloguing and therefore should be regarded as a landmark in the development of cataloguing. The most important events in the area of cataloguing and bibliographic description for this period are listed below chronologically:

1841 Panizzi's 91 rules

1852 Charles Coffin Jewett's code

1853 Charles Coffin Jewett's code (2nd ed.)

1867 Rules for cataloguing in congressional library

1876 Charles Ami Cutter's rules for a printed dictionary catalogue

1883 ALA's condensed rules for an author and title catalogue

1889 Cutter's rules for a dictionary catalogue (2nd ed.)

1891 Cutter's rules for a dictionary catalogue (3rd ed.)

1902 ALA's condensed rules for an author and title catalogue (Advanced ed.)

1904 Cutter's rules for a dictionary catalogue (4th ed.)

1905 Library of Congress supplementary rules on cataloguing

1906 Library of Congress special rules on cataloguing

1908 ALA and BLA's catalogue rules: author and title entries

1927 Vatican code

1931 Ranganathan's classified catalogue code

1931 Pierson's guide to the cataloguing of serials publications

1949 Rules for descriptive cataloguing in the Library of Congress

1961 Paris Principles

1967 AACR I (North American and British Text)

1968 MARC

1971 ISBD

1978 AACR II

During this period the much-required agreed-to general format for bibliographic description could not be developed. IFLA first initiated the development of such a general format in the form of ISBD(G) under its UBC programme.

NOTES

1.3 STANDARDS FOR BIBLIOGRAPHIC RECORD FORMAT

Standardisation of the record format implies the standardisation of the above mentioned three aspects at national, regional and international levels. Design and implementation of a standard record format ensure uniformity which is acceptable to all bibliographic agencies involved in information transfer which is very essential.

Background

Standardisation of the record format in manually prepared bibliographic lists started to be a matter of international concern from 1960s. The International Conference on Cataloguing Principles (ICCP) held in Paris in 1961 set up the standards for the heading of the author and title records in catalogues and bibliographies. The conference was sponsored by IFLA with the intention of evolving a set of basic principles to serve as guidelines in the design of catalogue codes all over the world. Paris principle could make some impact on certain national codes. However, differences in heading continued to exist in various catalogues and bibliographies and they stood in the way of interchange of information. The major effort for standardisation of record formats started from the development of ISBD.

International Standard Bibliographic Description (ISBD)

Starting from the first ISBD on monographs, a number of ISBDs have been developed including ISBD(G).

In 1973 the ISBD(M) text had been adopted by a number of national bibliographies and, translations of the original English text into several other languages had been done. By then it was realised that the printed word is not the only means of documentary transmission through which the communication needs of individuals and institutions are met. And that there was need for a standardised descriptive structure for documentary materials other than books. Consequently, the ISBD (NBM) International Standard Bibliographic Description for Non-Book Materials was published in 1977.

"This ISBD contained provisions covering machine-readable data files. However, when the ISBD(NBM) was being reviewed, together with the ISBD(CM), ISBD(M), and ISBD(S), by the ISBD Review Committee formed by IFLA in 1981, it was decided that special consideration should be given to the rapidly increasing need for a separate ISBD for computer files [ISBD(ER), 1997]." With the development of programs and data files for smaller computers, the nature of the medium became more complex; in addition, this change resulted in physical items roughly comparable to other library materials to be more widely added to library collections. Hence, bibliographic control was needed for them. As a result, the ISBD(CF) Working Group was established in 1986. In 1990, the first edition of ISBD(CF) was formally brought out.

With the emergence of interactive multimedia, development of optical technology availability of remote electronic resources on the Internet and World Wide Web, and reproductions of electronic resources, it was felt that ISBD(CF) should address the bibliographic implication of such developments. A Working Group was formed in 1994. In 1995, the Second Edition of the draft was prepared and distributed for worldwide review from individual readers, library associations and national libraries. As a result, many improvements were made, including recognition of the need for a new term to characterise the material under discussion. Thus, the more appropriate term 'electronic resource' was chosen.

Purpose of ISBD

The primary purpose of the ISBDs is to provide the stipulations for compatible descriptive cataloguing worldwide in order to aid the international exchange of bibliographic records between national bibliographic agencies and throughout the international library and information community. By specifying the elements which comprise a bibliographic description and by prescribing the order in which those elements should be presented and the punctuation by which they should be demarcated, the ISBDs aim to:

NOTES

- make records from different sources interchangeable, so that records produced in one country can be easily accepted in library catalogues or other bibliographic lists in any other country;
- assist in the interpretation of records across language barriers, so that records produced for users of one language can be interpreted by users of other languages; and
- assist in the conversion of bibliographic records to electronic form.

Machine-Readable Record Format

Library of Congress (LC) was the first to design and experiment on a MAchine-Readable Catalogue (MARC) record format for the purpose of communicating bibliographic information to large number of libraries. When MARC-I commenced as a pilot project in 1966 in LC, there were no established MARC formats available. Library professionals had reached no consensus as to what all access points were required for taking full advantage of an automated catalogue.

The pilot project known as MARC-I began in the year 1965 with the main aim of creation and distribution of machine-readable cataloging data to other libraries with Library of Congress (LC) as the distributing point. MARC-I only dealt with books. The development of MARC-II started in 1968. It was planned to cover all types of materials including books and monographs. During 1970–1973 documentation was issued for other materials, *i.e.*, in 1972 films records were issued, 1973 for serials, maps and French books and by 1975 records for German, Spanish, and Portuguese material [*Simmons and Hopkinson, 1988*].

In the year 1999, USMARC and CAN/MARC were harmonized and named as MARC21 [*McCallum 1989*]. The MARC21 bibliographic format, as well as all official MARC21 documentation, is maintained by the Library of Congress and by Canadian National Library [MARBI, 1996]. Recently UKMARC is also being merged with MARC21 and British Library is shifting from UKMARC to MARC21.

The Library of Congress and the National Library of Canada serve as the maintenance agency for the MARC21 formats for bibliographic, authority, holdings, classification, and community information data.

MARC Format

A MARC record involves three elements: the record structure, the content designation, and the data content of the record [MARBI, 1996]:

- *Structure:* MARC records is typical of Information Interchange Format (ANSI Z39.2) and Format for Information Exchange (ISO 2709).
- *Content designators:* By definition "the codes and conventions established to identify explicitly and characterise further the

NOTES

data elements within a record and to support the manipulation of those data". Anything that establishes the kind of data is a Content Designator, for example, there are three kinds of content designators

- tags, indicators, and subfield codes.

- *Content:* This is the actual data that is stored in the data fields. Often most of the data elements are defined by standards outside the formats. For example, Anglo-American Cataloguing Rules, Library of Congress Subject Headings, National Library of Medicine Classification.

  In MARC21, formats are defined for five types of data: Bibliographic, holdings, authority, classification, and community information.

MARC Record Structure

A typical MARC record consists of three main sections: the leader, the directory, and the variable fields [MARBI, 1996].

- The leader consists of data elements that contain coded values and are identified by relative character position. It is also called as Record label in CCF and UNIMARC. Data elements in this section define parameters for processing the record. It is fixed in length (24 characters) and occurs at the beginning of each MARC record.
- The directory contains the tag, starting location, and length of each field within the record. The length of the directory entry is defined in the entry map elements in Leader/20–23. In the MARC21 formats, the length of a directory entry is 12 characters, while in CCF it is 14 characters where character 13^{th} and 14^{th} are Segment Identifier and Occurrence Identifier. The directory ends with a field terminator character.
- The data content of a record is divided into variable fields. The MARC21 format distinguishes two types of variable fields: variable control fields and variable data fields.

UNIMARC

With a view to facilitate international exchange of MARC data, IFLA established a Working Group on Content Designators in 1972. The Working Group was charged with analysing the differences in content designation among the national formats that had been developed to that point in time, exploring ways of accommodating those differences, and recommending a uniform set of content designators that would serve as a standard for international exchange [IFLA, 1977].

NOTES

The format developed by the Working Group became UNIMARC. UNIMARC was conceived not as a standard to be imposed on bibliographic agencies as a national communications format, but rather as a 'translation' mechanism to be used by those agencies when exchanging data across national borders. The idea was that UNIMARC would serve as the common 'vocabulary' that would function as a means of conveying data that had originally been encoded in the national format of the sender in a commonly recognised form that could subsequently be converted into the national format of the receiver. Each national bibliographic agency would therefore have to develop and maintain only one conversion program to convert from their national format to UNIMARC and one to convert from UNIMARC to the national format, rather than multiple programs to convert from one national format to another on a one-to-one basis.

The purpose of UNIMARC is to facilitate the description, retrieval and control of bibliographic items for data exchange and also for local bibliographic format. There are a number of factors that shaped UNIMARC into the flexible data package [*Morataza, 1996*].

Block Structure: One key objective of the format is to be able to accept easily the data from a number of different national formats. That is the basic requirement behind the block structure of UNIMARC. Since data may be carried in a number of different positions in various national formats, the emphasis in designing UNIMARC was to identify functionally the different types of data and establish clearly designated areas for them. On receipt of UNIMARC records, national systems may rearrange data in any way that is practical at the local level. The UNIMARC areas are called Blocks, and the block number forms the first digit of the tags for data fields in the block.

Record Structure: It required supporting International Standard Exchange Format, hence it adheres to ISO 2709 for record structure.

ISBD: Another important point for UNIMARC was appropriate support of the international description guidelines that were developed by IFLA and incorporated into cataloguing codes worldwide. The ISBD data is especially accommodated in UNIMARC Block 2, the descriptive paragraph.

Textual and Non-textual Material: Fourth requirement of UNIMARC is that it should be able to accommodate description of wide variety of materials. Hence, it covers books, periodicals, maps, globes, music scores, sound recordings, motion pictures, video recordings, pictures, drawings, sculpture, artifacts, computer files and other related items.

Multiple Levels: Another basic requirement for UNIMARC is that it should accommodate various bibliographic levels, not just monographic (single part) material. Serially issued items (multiple part items) and analytic (part of larger item) need to be included, and these levels can occur for any of the above types of materials. Again the flexibility

NOTES

of the coded data field is important. Serial related fields are not confined to use in records for journals or the traditional forms of serially issued items, but can be included in records for serially issued maps, films, etc.

Parts of Items/Linking Technique: The provision of describing parts of items (*e.g.*, journal articles, chapters) in UNIMARC follows the common practice in the library to provide a citation to the host (book or serial that contains the part) but provide a separate record to fully describe the host bibliographically. UNIMARC thus focuses on an analytical record on the part being catalogued – the monograph in series, journal article, books part, etc. Only linking entry field gives information about the host item, such as series (for a monograph in a series), the serial (for a journal article).

UNIMARC Functional Blocks

The fields of the bibliographic record are divided into functional blocks, the first digit of the tag indicates the block of the field [*McCallum, 1989*].

1. Identification block, contains those numbers that identify the record of the work (*e.g.*, ISBN, ISSN).
2. Coded Information block, contains fixed length coded data elements describing various aspects of the work.
3. Descriptive block, contains the areas covered by the ISBD (*i.e.*, title, edition, imprint, collation, series) with the exception of standard number and notes.
4. Notes block, contains free text statements describing various aspects of the work.
5. Linking Entry block, contains standard links in numeric and textual form to other records.
6. Related Title block,contains titles to be used as access points.
7. Subject Analysis block, contains subject identification (*e.g.*, UDC, Library of Congress Subject Headings, etc.). Personal and corporate names used as subjects will appear in this block.
8. Intellectual Responsibility block, contains names of persons and corporate bodies responsible for the creation of the work described in access point form.
9. International use block, contains internationally agreed field that do not fit in the preceding blocks, 1. to 8.. This block includes fields on originating agency, ISSN Centre, general cataloguer's note and electronic location and access.
10. National Use block, reserved for national use by agencies where UNIMARC is the basis of the domestic format.

NOTES

Common Communication Format (CCF)

UNESCO was concerned in the late 1970s about the lack of developments in the field of scientific information, especially in systems for the sharing of information on journal articles. Especially in the light that secondary services such as abstracting and information services were being automated. UNESCO sponsored an international symposium in 1978 to look at the problems caused by this sector having many different formats and recommended developing a switching format taking into account the need to be compatible with MARC via UNIMARC, and the secondary services. Those experts who had developed ISO 2709 also attended the meeting since the format required more sophisticated linking features in order to be able to link records of articles and the journals and issues containing them. The symposium set-up the UNESCO Ad-hoc Group on a Common Communication Format and the CCF was developed. This was mainly used by the secondary services in the science and technological sectors by UN agencies and more generally in India [*Hopkinson, 1989*].

It is a specific implementation of the ISO 2709. The record label consists of 24 characters and directory of 14 characters. Data field consist of indicators, subfields and data field separator. A CCF record may contain descriptions of more than one item, but the description of each item occupies a single record segment. The major item occupies the primary segment and the others the secondary segments. Segment links are used in vertical relationships (*e.g.*, a monograph and a unit in it) and horizontal relationships (*e.g.*, versions of a work in different languages) are also used. For details about the CCF record structure and data elements.

Indian Standard

Standardisation of record format has not received due attention in Indian libraries. At national level Indian Standards Institution [renamed as Bureau of Indian Standards (BIS)] had evolved a standard for bibliographic references in 1963 for use in non-computerised systems. However, it could not keep track of the developments in the media and forms of documents. In the following years a revision of the standard appeared in 1979 wherein ISBD was suggested as a substitute format to be adopted by agencies willing to do so.

The sixteenth Indian Standards Convention of ISI held in Bhopal in October 1975 discussed the issue of standardisation of bibliographic information in the context of machine-readable records. Continued efforts in this area by ISI resulted in the design and publication of a standard (IS:11370–1985) titled, 'Guide for data elements and record format for computer based bibliographic description of different kind of documents' in July 1986. Considerable assistance in the preparation of this standard has been taken from international standards relevant

to machine-readable record format. Structure of this format conforms to ISO 2709: 1981.

Among the data elements, descriptive block is based on ISBD. Subject analysis block includes POPSI, PRECIS, keyword and synopsis or abstract. There is a local use block which is intended to communicate the variations made in the existing data and for the inclusion of new data needed for local use.

The standard specifies the requirements for machine-readable record format for books, periodicals, conference proceedings, articles in a periodical, research reports of ongoing research projects, patents and standards.

However, the latest decision of National Library of India is to follow MARC21 standard and retro-conversion work is already in progress.

1.4 BIBLIOGRAPHIC DESCRIPTION OF NON-PRINT MATERIALS

The non-book materials (NBM) have special features which are different from books. Many non-book materials are machine dependent meaning thereby for consultation of the material or identification of sources of information for the description, equipments are required. The description of such non-book materials requires full description of media, creators and other elements. In contrast, for books many of the information can be had by directly browsing the books on the shelves. The non-book materials are available in varieties of formats and media, such as:

Medium	***Format***
Films	Filmstrip
Magnetic Tapes	Slides
Electronic Media	Cine films
Microforms	Microcards
Sound Tapes	Microfiches
Videos	Micro opaque
Plastic Materials	Microfilms
Cartographic Materials	Video cassettes
	Videotapes
	Audio cassettes
	Audiotapes
	Paintings
	Charts
	Records

The bibliographic description of NBM are similar to books and other materials. The main structure of an entry comprises of heading and

NOTES

body of the entry (description). The major area which creates problem in identification of different elements of description of NBM are author, title, physical description and subject. The sources of information for these elements are the material itself, accompanying text and other sources such as informal guides. An example of bibliographic description for a non-book material is given below:

Table 1: AACR2: General Rules for the Description of Library Materials, together with a Worked-out Example

1.1	Title and statement of responsibility area	
	B. Title Proper	The librarian
	C. General material designation	(graphic)
	E. Other title information	Personality plus
	F. Statement of responsibility	compile by Jack Lurcher Photograph by Susan Shera
1.2	Edition area	
	B. Edition statement	2nd Ed.
1.3	Material (or type of application) specific area. No general use of this areas is envisaged for NBM. However, if an item is being described whose contents fall within the scope of cartographic materials, serials, publications, music, computer files and in some circumstances, microfilms.	
1.4	Publication, distribution, etc., area	
	C. Place of publication, distribution	Newcastle Luton
	D. Name of publisher, distributors	Rectory Publications Bishopscotes
	E. Statement of function of publisher, distributor, etc.	[production company] [distributor]
	F. Date of publication, distribution, etc.	1988
	G. Place of manufacture, name of manufacturer, date of manufacture, etc.	
1.5	Physical description area	
	B. Extent of item (including specific material designation)	36 slides
	C. Other physical details	Col
	D. Dimensions	
	E. Accompanying material	+ 1 booklet (18p; 16 cm)

NOTES

1.6	Series area	
	B. Title proper of series	(Media and the
	G. Numbering within series	librarian 5)
1.7	Note area	
	B. Notes	Also available in filmstrip version. Illustrates the vital role of librarian in encouraging use of NBM.
1.8	Standard number and terms of availability area	
	B. Standard number	0-85365-509X
	D. Terms of availability	£35.00

Guidelines for Bibliographic Description of Electronic Resources

The first edition of International Standard Bibliographic Description for Non-Book Materials [ISBD(NBM)] was produced and published in 1977. This ISBD(NBM) contained provisions covering machine-readable data files. The ISBD Review Committee, formed by IFLA in 1981, decided to give special consideration to cope with the rapidly increasing need for a separate ISBD for computer files. As a result, the ISBD(CF) was published. Electronic resources are products of a volatile technology that continues to generate changes at a very rapid pace. Specific among recent advances are the following: emergence of interactive multimedia; development of optical technology; availability of remote electronic resources on the Internet and World Wide Web; and reproductions of electronic resources. In addition, many improvements are realised, including recognition of the need for a new term 'electronic resource' which is judged more appropriate than the term 'computer file' used previously. Hence, the ISBD(ER) is originated during 1999–2000.

ISBD(ER)

The International Standard Bibliographic Description for Electronic Resources [ISBD(ER)] is one of several published ISBDs to specify the requirements for the description and identification of such items, to assign an order to the elements of the description, and to specify further a system of punctuation for the description. Electronic resources consist of materials that are computer-controlled, including materials that require the use of a peripheral (*e.g.*, a CD-ROM/DVD player) attached to a computer; the items may or may not be used in the interactive mode. Included are two types of resources: data (information in the form of numbers, letters, graphics, images, and sound, or a combination thereof) and programs (instructions or routines for performing certain tasks including the processing of data). In addition, they may

NOTES

be combined to include electronic data and programs (*e.g.*, online services, interactive multimedia). For cataloguing purposes, electronic resources are treated in the ISBD(ER) in two ways depending on whether access is local or remote. Local access is understood to mean that a physical carrier can be described. Such a carrier (*e.g.*, disk/ disc, cassette, cartridge, etc.) must be inserted by the user into a computer or into a peripheral attached to a computer - typically a microcomputer. Some other devices, like, Personal Digital Assistant (PDA), VCD/DVD player along with television set, e-Book Reader, etc., can also provide access to electronic information resources. Remote access is understood to mean that no physical carrier can be handled by the user - typically, access can only be provided by use of an input-output device (*e.g.*, a terminal) either connected to a computer system (*e.g.*, a resource in a network) or by use of resources stored in a hard disk or other storage device. However, in cases where electronic resources combine the characteristics described in other ISBDs (*e.g.*, an electronic serial, digitised map, etc.), it is recommended that the bibliographic agency first make full use of the stipulations in the ISBD(ER) and apply provisions of other ISBDs as appropriate, if necessary.

Prescribed Sources of Information

The bibliographic description for Electronic Resources is completed in eight areas. Information is collected from chief source of information which are prescribed as below.

Area	Prescribed Sources of Information
1. Title and statement of responsibility	Internal sources; labels on the physical carrier; documentation, containers, or other accompanying material
2. Edition	Internal sources; labels on the physical carrier; documentation, containers, or other accompanying material
3. Type and extent of resource	Any source
4. Publication, distribution, etc.	Internal sources; labels on the physical carrier; documentation, containers, or other accompanying material
5. Physical description	Any source
6. Series	Internal sources; labels on the physical carrier; documentation, containers, or other accompanying material
7. Note	Any source
8. Standard number (or alternative) and terms of availability	Any source

NOTES

Information taken from outside the prescribed source(s) is tc be enclosed in square brackets.

1.5 SUMMARY

- Libraries are no longer individual systems that can operate in isolation. There is a trend towards universal access to library data. Hence it is very important to maintain international standards for library operations especially for bibliographic standards. This unit has covered in detail the different standards for bibliographic record format.
- International standards laid down by IFLA and other agencies, ISBD description has been described. In machine-readable record (MARC) format, MARC21 that is fast becoming the de-facto standard, has been dealt in detail.
- Common Communication Format for the exchange of bibliographic records and structure of CCF record is explained.
- Indian libraries and documentation centres had been passive towards standardisation of bibliographic record format in the past. Lack of communication facilities, non-availability of bibliographies in machine-readable form, diversity of languages and slow pace of computerisation might be the possible reason for this. Moreover, lack of awareness of the significance of the standardisation is pointed out by ISO as the major obstacle in this path. Of late the major libraries including National Library of India are adopting MARC21 standard for bibliographic records thus following international trends.

1.6 REVIEW QUESTIONS

1. Write short note on Record Format.
2. Distinguish between fixed field and variable field in a record.
3. How ISBD helps in the standardisation of bibliographic records?
4. Explain the significance of Common Communication Format (CCF).
5. Explain the multidimensional scope of bibliographic description in online environment.

1.7 FURTHER READINGS

1. Fothergill, Richard and Butchart, Ian (1990). *Non-book Material in Libraries: A practice Guide.* 31" ed. London: Clive Bingley.

2. Graham, Paul (1985). Current Developments in Audiovisual Cataloguing. *Library Trends, Summer,* pp 5–66.

3. IGNOU, (1994). *Bibliographic Description of Non-Print Materials.* MLIs-03, Block 2, Unit. pp 4–52.

4. Rogers, Jo Ann V. and Saye, Jerry D. (1987). *Non-Print Cataloguing for Multimedia Collections: A Guide based on AACR 2.* 2nd ed. Colorado: Library Unlimited.

5. Wall, Thomas B. (1985). Non-Print Materials: A Definition and some Practical Consideration of their Maintenance. *Library Trends,* Summer 1985, pp 129–40.

6. Weihs, Jean [*et al.,* (1979)]. *Non Book Materials: The Organic Section of Integrated Collection.* 2' ed. Canada: Canadian Library Association.

NOTES

NOTES

UNIT II SUBJECT ANALYSIS AND INDEXING

★ STRUCTURE ★

LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain indexing principles and process
- know about comparison between controlled and natural language indexing
- describe the Universal Decimal Classification (UDC) scheme
- explain subject indexing systems
- know about automatic indexing
- describe the machine translation.

2.1 INTRODUCTION

An index is a guide to the items contained in or concepts derived from a collection. Item denotes any book, article, report, abstract review etc., (textbook, part of a collection, passage in a book, an article in a journal etc.). The word index has its origin in Latin and means: 'to point out, to guide, to direct, to locate'. An index indicates or refers to the location of an object or idea. The definition according to the British Standards (BS 3700: 1964) is "a systematic guide to the text of any reading matter or to the contents of other collected documentary

material, comprising a series of entries, with headings arranged in alphabetical or other chosen order and with references to show where each item indexed is located". An index is, thus, a working tool designed to help the user to find his way out mass of documented information in a given subject field, or document store. It gives subject access to documents irrespective their physical forms like books, periodical articles, newspapers, AV documents, and computer-readable records including web resources.

Early indexes were limited to personal names or occurrences of words in the text indexed, rather than topical (subject concept) indexes. Topical indexes are found the beginning of the 18th century. In the nineteenth century, subject access to books was by means of a classification. Books were arranged by subject and their surrogates were correspondingly arranged in a classified catalogue. Only in the late 19th century, subject indexing became widespread and more systematic. Preparation of back-of-the-book index, historically, may be regarded as the father of all indexing techniques. Indexing techniques actually originated from these indexes. It was of two types: Specific index, which shows broad topic on the form of one-idea-one-entry, *i.e.*, specific context of a specific idea; and Relative index, which shows various aspects of an idea and its relationship with other ideas. Specific index cannot show this, it only shows broad topic on the form of one-idea-one-entry, *i.e.*, specific context of a specific idea. The readymade lists of subject headings like Sears List and LCSH fall far short of actual requirement for depth indexing of micro documents in the sense that the terms are found to be too broad in the context of users' areas of interest and of the thought content of present day micro document.

2.2 CLASSIFICATION SCHEMES

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

Classification schemes belong to different categories, viz. enumerative, analytico synthetic and faceted. The degrees of enumerativeness and facetedness vary from partial to full. Library of Congress scheme

belongs to the former category and Colon Classification to the latter. Dewey Decimal Classification (DDC) began as an enumerative scheme but now it has incorporated quite a good degree of facetedness. Subjects of different kinds varying from simple to complex are all enumerated along with their notation in an enumerative scheme. In a faceted scheme only the possible facets are enumerated, a subject can only be represented by joining the facets after analysis of the subject.

The Universal Decimal Classification scheme is an example of analytic synthetic scheme which is not freely faceted.

2.3 UNIVERSAL DECIMAL CLASSIFICATION (UDC) SCHEME

The Universal Decimal Classification (UDC) is a worldwide popular general classification scheme covering all fields of knowledge. It is an analytico-synthetic classification scheme, where the concepts can be broken down into simple concepts (analysed) and then combined (synthesised). The UDC can be applied in different contexts that make it more flexible than other general classification schemes. It is considered as a sophisticated indexing and retrieval tool for documentation and information services.

Historical Background

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

The UDC Consortium

Presently, UDC is owned by the UDC Consortium (UDCC) since January 1992. The UDCC maintains multilingual Master Reference File

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

English Editions of UDC

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

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

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

(*iii*) **The UDC Abridged Edition.** The latest UDC Abridged Edition, published in 2003, contains over 4,100 entries. It was formerly known as the UDC pocket edition.

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

Features and Structure of the UDC

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

NOTES

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

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

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

Notation

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

Structure of UDC (BS1000M: 1993)

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

Main Tables

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

1. Generalities
2. Philosophy. Psychology
3. Religion: Theology
4. Social Sciences
5. (kept vacant at present)
6. Mathematics and Natural Sciences
7. Applied Sciences
8. Fine arts. Applied arts. Entertainment. Games. Sports
9. Language. Linguistics. Literature
10. Archeology. Geography. Biography. History

Each main class is subdivided into its logical parts, with each subdivision further subdivided. The more detailed the subdivision, the longer the number that represents it. For ease of reading, long notational elements are broken up into 3-digit units by means of the point (.), which has no other significance (*e.g.*, 123.4 or 123.456.7 and so on).

But unlike DDC, UDC does not use 0 (zero) in the last digits for the first two levels of main classes (*e.g.*, 5, 51; instead of 500, 510 as followed in DDC). An example is shown below to give you an idea on the hierarchical structure of UDC, as described in the Main Tables:

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

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

Auxiliary Tables

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

Common Auxiliary Tables

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

Table Ia Co-ordination Addition

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

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

(540 + 510) India and China

51 + 53 Mathematics and Physics

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

(4/6) A group of Europe, Asia and Africa

592/599 Systematic Zoology (class 592 to 5999)

Table Ib Relation. Sub-grouping. Order-fixing

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

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

316.64:342 Social view in public law

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

7:17 Art in relation to ethics

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

77.044::796 Sports photography

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

783:[283/289] Protestant Church Music

Table Ic Common Auxiliaries of Language

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

=214.21 Hindi

51=214.21 Mathematics in Hindi language

Table Id Common Auxiliaries of Form

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

(031) Encyclopedia

5/6(031) Encyclopedia of Science and Technology

Table Ie Common Auxiliaries of Place

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

(1–11) Eastern

(261.268) English Channel

(541.23) West Bengal

37(541.23) Education in West Bengal

Table If Common Auxiliaries of Ethnic Grouping and Nationality

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

(=1.540) Indian

(=1.4) European

(=1.253) Forest dwellers

Table Ig Common Auxiliaries of Time

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

"20" Twenty first century

5"20" Science in the twenty first century

"3453" Evening

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

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

523.44*433 Minor planet Eros (IAU authorised number for Eros is 433)

625.711(540)*NH8 Road engineering with special reference to Indian national highway number 8

Table Ii Common Auxiliaries of Point-of-view

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

0.003 Economic point of view

6.003 Economic point of view of Technologies and Applied Sciences

0.002.6 Product point of view

688.322.002.6 Digital computer as a product

Table Ij Common Auxiliaries of Materials, Persons and Personal Characteristics.

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

–033.52 Crystal glass

NOTES

730–033.52	Sculpture using crystal glass
–035.54	Furs

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

–058.856	Foster parents
–057.81	Illiterates
730–57.81	Sculpture, made by illiterate persons

Special Auxiliary Subdivisions

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

The special auxiliary subdivisions use three kinds of notation

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

Example:	82–1/–9	English Literary forms, genres
	82–1	English Poetry. Poems. Verse

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

Example:	528	Geodesy. Surveying. Photogrammetry. Cartography
	528.01	Preparatory work. Making of stations. Signal construction.

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

Filing Order

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

Citation Order

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

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

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

Qualities of the UDC Scheme

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

(*i*) The UDC is very flexible in nature.

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

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

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

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

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

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

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

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

NOTES

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

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

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

The analytico-synthetic nature of UDC scheme makes it very flexible and accommodating. Although this scheme is not fully faceted, it has very powerful facility, which permits the joining of any part of the classification with any other.

UDC can be applied in different contexts and in different applications in special libraries and information centres. It is now being vigorously geared for use in online catalogues and websites on the World Wide Web.

2.4 INDEXING PRINCIPLES AND PROCESS

Purpose of Indexing

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

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

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

NOTES

Problems in Indexing

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

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

(*b*) Multidimensional users need for information;

(*c*) Choice of terms from several synonyms;

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

(*e*) Distinguishing homographs;

(*f*) Identifying term relationships—Syntactic and Semantic;

(*g*) Depth of indexing (Exhaustivity);

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

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

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

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

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

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

When indexing, it is necessary to understand, at least in general terms, what the document is about (aboutness). The subject content of a document comprises a number of concepts or ideas. For *e.g.*, an article on lubricants for cold rolling of aluminium alloys will contain information on lubricants, cold rolling, aluminium alloys etc. The indexer

selects these concepts, which are of potential value for the purpose of retrieval, *i.e.*, those concepts on which according to him, information is likely to be sought for by the users. It is the choice of concepts or the inner ability to recognise what a document is about is in the very heart of the indexing procedure. However, it is the identification of concepts that contributes to inconsistencies in indexing.

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

Indexing Process

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

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

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

(*c*) What is the nature of collection?

(*d*) What is the characteristics of user population?

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

(*f*) Display or physical appearance of the index.

Essentially, the processes of indexing consist of two stages:

(*i*) establishing the concepts expressed in a document, *i.e.*, the subject; and

(*ii*) translating these concepts into the components of the indexing language.

(*a*) **Establishing the Concepts Expressed in a Document.**

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

(*i*) **Understanding the Concepts.** Full comprehension about the content of the documents depends to a large extent on the form of the document. Two different cases can be distinguished, *i.e.*, printed documents and non-printed documents. Full understanding of the printed documents depends upon an extensive reading of the text. However, this is not usually practicable, nor is it always necessary. The important parts of the text need to be considered carefully

with particular attention to: title, abstract, introduction, the opening phrases of chapters and paragraphs, illustrations, tables, diagrams and their captions, the conclusion, words or groups of words which are underlined or printed in an unusual typeface. The author's intentions are usually stated in the introductory sections, while the final sections generally state how far these aims are achieved. The indexer should scan all these elements during his study of the document. Indexing directly from the title is not recommended, and an abstract, if available should not be regarded as a satisfactory substitute for a reading of the text. Titles may be misleading; both titles and abstracts may be inadequate in many cases, neither is a reliable source of the kind of information required by an indexer.

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

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

(*iii*) **Selection of Concepts.** The indexer does not necessarily need to retain, as indexing elements, all the concepts identified during the examination of the document. The choice of those concepts, which should be selected or rejected, depends on the purpose for which the indexing data will be used. Various kinds of purpose can be identified, ranging from the production of printed alphabetical indexes to the mechanized storage of data elements for subsequent retrieval. The kind of document being indexed may also affect the product. For example, indexing derived directly from the text of books, journal articles, etc., is likely to differ from that derived only from abstracts.

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

Indexing Language

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

Theory of Indexing

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

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

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

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

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

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

(*e*) **Inferential Level:** An indexer is able to make inferences about the relationships between words or phrases by observing the paragraph and sentence structure, and stripping the sentence of extraneous detail. This inference level makes it possible for the indexer to identify novel subject areas.

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

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

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

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

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

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

NOTES

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

Indexing Criteria

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

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

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

Indexing Policy: Exhaustivity and Specificity

Exhaustivity is a matter of an indexing policy and it is the measure of the extent to which all the distinct subjects are discussed in a particular document are recognized in indexing operation, and translated into the language of the system. Exhaustivity in indexing requires more number of index entries focusing different concepts (both primary and secondary) covered in the documents. The greater the number of

concepts selected for indexing purpose, the more exhaustive is the indexing. If, in a given document, concepts A, B, C, D, E are selected for indexing then the indexing of the document is more exhaustive than if only concepts A< B< C are selected. When a relatively large number of concepts are indexed for each document, the policy followed is one of depth of indexing. Depth of indexing, in other words, allows for the recognition of concepts embodied not only in the main theme of the document but also in sub-themes of varying importance. Policy decision in respect of exhaustivity in indexing depends upon several factors like strength of collection, manpower available, economy and requirements of users.

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

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

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

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

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

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

NOTES

fully developed, or is referred to only casually by the author, indexing at a more general level may be justified.

Both Exhaustivity and Specificity are very closely related to recall and precision.

A high level of exhaustivity increases recall and high level of specificity increases Precision.

Quality Control in Indexing

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

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

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

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

Translation failures may be of three types:

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

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

(*c*) Omission of important term(s).

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

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

NOTES

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

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

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

Types of Indexing

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

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

Indexing Language

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

From now on we shall use the term language to refer to only the language that is used by humans. The characteristics of a language are vocabulary and rules for their arrangement (syntax). The languages may be artificial and natural. Natural languages refer to our languages, which we normally use for communication, whereas, artificial are those that we have designed for a specific purpose or are used in a specific sense or for limited use only. Shorthand is an example of this category which all of us have heard about. Similarly, we have examples of artificial languages in different disciplines *e.g.*, in chemistry we have a language to indicate names of different elements, compounds and also the process of their transformation. Similarly, notation of a classification scheme is an artificial language.

Types

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

Vocabulary Control

Indexing may be thought of as a process of labelling items for future reference. Considerable order can be introduced into the process by standardising the terms that are to be used as labels. This standardisation is known as vocabulary control, the systematic selection of preferred terms.

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

NOTES

step involves translation of the conceptual analysis to the vocabulary of the system. Thus, there is a close resemblance between indexing and search process.

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

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

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

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

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

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

From the earlier discussion it should be clear that a natural language system suffers from varieties of problems in the context of development of an index file. Thus, the need for control of the vocabularies arises. A controlled vocabulary refers to an authority list of terms showing the interrelationships and indicating the ways in which they may be combined to represent specific subject of a document. A certain degree of semantic structure is introduced in the controlled vocabulary so that

terms whose meanings are related may be brought together or linked in some ways. This semantic structure is incorporated by means of (*a*) controlling the synonyms, word forms, etc., and distinguishing homographs for consistent representation of the subject of the documents; and (*b*) providing mechanism to link the hierarchical and non-hierarchical terms that are related semantically to facilitate comprehensive search. Different techniques of vocabulary control have been adopted in the tools have List of Subject Headings (LSH), Thesaurus, Thesaurofacet, etc.

Controlled vs. Natural Language Indexing

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

Table 2.1: Comparison between Controlled and Natural Language Indexing

Natural Language	Controlled Language
High specificity gives precision; excels in retrieving individual terms — names of persons, organisations, etc.	Lacks specificity, even in detailed systems
Exhaustivity gives potential for high recall	Lacks exhaustivity; cost of indexing to the level of natural language is high; terms may be omitted due to indexer errors
Up-to-date; new terms are immediately available	Not possible to make it up-to-date; new terms are added to the thesaurus only at certain intervals
Words used by the authors are used in indexing; there is no provision for misinterpretation	Indexers may misinterpret some words; there is a risk of loss of terms and/or their proper meaning/use
Natural language used by the searcher is used for matching	The searcher may have to use artificial language (the controlled vocabulary)
Input cost is low since there is no human intervention	Input cost is high since human intervention may be required for translation of terms into appropriate controlled vocabulary terms
Language incompatibility can be overcome due to the use of natural language, and thus data exchange and cross-database searching becomes easier	The databases will have to use the same vocabulary control tool in order to be compatible
Intellectual effort is placed on searcher. Problems arise with	Eases the burden of searching: – controls synonyms and near synonyms

terms having many synonyms and near-synonyms	and leads to specific preferred terms to broaden search – qualifies homographs – provides scope notes – displays broader, narrower and related terms – expresses concepts elusive in free text
Syntax problems: there is a danger of false drops through incorrect term association	Overcomes syntax problems with compound terms and other devices
Exhaustivity may lead to loss of precision	Loss of precision can be avoided by selecting specific/narrower terms while searching

2.5 SUBJECT INDEXING SYSTEMS

You have learnt that one of the most important functions of a library catalogue is to provide access to documents in a library through their subject contents. A classified catalogue facilitates subject approach to documents in a library by subject arrangement of document surrogates in an organised and systematic way by using a system of classification chosen for the library. Nevertheless, an alphabetical subject index to the classified part of the catalogue is essential as the notation of the classification system (class numbers), by which the main entries of documents are arranged, is not obvious to most readers. A dictionary catalogue, however, provides this facility of subject approach through verbal subject representation of the content of documents. It should, therefore, be clear that irrespective of the inner forms of library catalogues, verbal subject headings are essential to satisfy subject approach to documents in a library. Just as we have codes for descriptive cataloguing that provide rules and procedures with reference to the choice and rendering of bibliographical data of documents, indexing systems have been designed and developed with their own standard rules and procedures for constructing subject headings.

Need and Purposes

The need and purpose of providing standard rules and procedures for constructing subject headings for documents arises due to various problems, some of which are discussed below:

(*i*) If the thought content of documents could be represented adequately by a single word, there will be no difficulty at all in deriving subject headings. For example, documents dealing with concepts such as, teaching, schools, industries, skating, can be easily

NOTES

represented as they are, but it is always not so. Several single concepts are multiworded; in which one word is the focus and the other(s) are qualifiers.

(*ii*) Another problem very much intrinsic to constructing subject headings is representing the concepts discussed in a document in their contextual relations. It is only in their combination that each concept could have any meaning with reference to the particular document(s). For example, consider the following:

Teaching chemistry in higher secondary schools using audio-visual aids.

Here, the concepts are: Teaching, Chemistry, Higher Secondary Schools, Audio-Visual Aids.

In this example, there are single words and multiple words representing the different concepts. All the concepts together represent the specific subject of the document. Here the word order is very important, as each concept is put in its right context to make it meaningful. This word order or the citation order of words is referred to in indexing parlance as *Syntactical Order or Syntax*. In other words, they are understood in their. syntactical relations context.

The problem in these types of examples, is to the fixed citation order of concepts and what should be done if the approach of readers is other than the one chosen. It is not possible to provide access points from all terms for economic and other reasons.

(*iii*) Yet, another problem is the intrinsic relation of concepts with other concepts. For example, teaching is related to learning, intelligence, assimilation, students and many other similar concepts. If a user is interested in teaching, there may be documents in the library that may deal with learning, intelligence and others that may have some relevance to the user. This type of relationship is referred to in indexing parlance as *Semantic* relations, The question is how are such relations taken care of in subject indexing?

In summary, the main purpose of prescribing standard rules and procedures for constructing subject headings is:

- to prescribe a standard methodology to be used by subject cataloguers and indexers for constructing subject headings;
- to be consistent in the choice and rendering of subject entries, using standard vocabulary and according to given rules and procedures for displaying syntactic and semantic relationships;

- to be helpful to readers in accessing any desired document(s) from the library catalogue through different ways of such an approach;
- to decide on the optimum number of subject entries, and thus economies on the bulk and cost of cataloguing.

In short, the main purpose is to provide maximum help to users of library catalogues by proper and consistent subject approach at optimal cost.

Subject heading systems provide specific guidelines to deal with the different problems mentioned above, on the basis of principles and postulates.

Factors Governing Subject Indexing Systems

There are many factors that govern the design of subject heading systems. Some of the more important ones are:

- content analysis of documents to select the right key words that represent their specific subject;
- rendering the selected key words in a logical sequence according to prescribed principles and postulates;
- establishing main and added entries in standard formats;
- provision of cross references to subject concepts to obtain as many relevant documents as are available in the library through catalogue or index file; and
- arrangement of entries.

Content analysis of documents is totally independent of any technique of indexing. Skills required in this process are the subject background of the indexer on the subject(s) indexed and the ability to read/scan documents fairly quickly for the purpose of identifying appropriate key words to represent the thought contents of documents.

To know what a document is about, some of the aids available in the documents themselves are: Titles and sub-titles, if any; content page(s) with all the details of chapters listed; foreword, preface; conspectus; introduction; the text — a quick scanning of the body of the text; and the index. Sometimes the captions in tables, diagrams, etc., and book jacket serve as an useful aid. In addition to all these aids, dictionaries, encyclopaedias; handbooks, guides etc., are useful consulting aids to know the subject(s). It may also be necessary on occasions to consult subject experts to know subject ramifications and their importance.

For the rest of the factors mentioned above, indexing systems usually provide the necessary guidelines, rules and procedures, backed up by principles, postulates, etc., which would explain the designer's logic and philosophy of approach. The understanding of all these and their

NOTES

proper assimilation by indexers would ensure a fairly good quality subject index.

Subject Indexing Models and their Principal Features

A number of subject indexing models have been designed and developed in the past hundred years or so, many of which are in use in libraries and other bibliographical publications all over the world. We can classify them to be of following four kinds:

(*a*) Authority lists based Subject Indexing Models, *e.g.*, Library of Congress Subject Headings (LCSH), Sears List of Subject Headings (SLSH), etc.

(*b*) Pre-coordinate Indexing Models, *e.g.*, Chain procedure, PRECIS, COMPASS, POPSI, etc.

(*c*) Post-coordinate Indexing Models, *e.g.*, Uniterm Indexing.

(*d*) Keyword-based Indexing Models, *e.g.*, KWIC, KWAC, KWOC, etc.

We shall study authority list-based subject indexing models, such as LCSH and SLSH, in terms of their back-ground, formation of subject headings in accordance with prescribed principles and postulates, entry format, cross references, filing order, standard vocabulary used, its updating mechanism, merits and deficiencies. We shall also discuss in brief the basic and principle features of subject indexing systems by Chain Procedure, PRECIS, COMPASS, POPSI, Uniterm, and Keyword indexing.

2.6 AUTOMATIC INDEXING

In many literatures of Library and Information Science, the term 'automatic indexing' is interchangeably used with the term 'computerised indexing'. A fully automatic indexing system would be one in which indexing is conducted by computers, an internally generated thesaurus is prepared, and search strategies are developed automatically from a natural language statement of information need. Salton provides the following definition of automatic indexing: When the assignment of the content identifier is carried out with the aid of modern computing equipment the operation becomes automatic indexing. It has been suggested that the subject of a document can be derived by a mechanical analysis of the words in a document and by their arrangement in a text. In fact, all attempts at automatic indexing depend in some way or other on the test of the original document or its surrogates. The words occurring in each document are examined and substantive words are selected through statistical measurements (like word frequency

NOTES

calculation, total collection frequency, or frequency distribution across the documents of the collection) by the computer.

However, the use of computers in generating indexes of documents started from KWIC indexing developed by H.P. Luhn.

The idea of analysing the subject of a document through automatic counting of term occurrences was first put forward by H.P. Luhn of IBM in 1957. He proposed that:

(*a*) The frequency of word occurrence in a text of the document furnishes a useful measure of word significance;

(*b*) The relative position of a word within sentence furnishes a useful measurement for determining the significance of sentences; and

(*c*) The significance factor of a sentence will be based on a combination of these two requirements.

The basic idea behind Luhn's automatic indexing was based on word extraction, that is, keywords were extracted from the text by counting the frequency of occurrence of words in a given document. Here, the computer was used to scan the text with the object of counting the words or phrases that occur most frequently in a machine-readable document, and the extraction programs select the words or phrases that occur most frequently to represent the subject-matter of the document. A 'stop word' list was first used to eliminate the common and non-substantive words. The system pioneered by Luhn was relatively effective and the words or phrases selected by computer were quite similar to those, which would be extracted by a human indexer.

In the early 1960s, some other attempts were made at implementing automatic indexing systems. These consisted in using the computer to scan document texts, or text excerpts such as abstracts, and in assigning as content descriptor words that occurred sufficiently frequently in a given text. A less common approach uses relative frequency in place of absolute frequency. In relative frequency approach, a word is extracted if it occurs more frequently than expected in a particular corpus. Thus in a document on 'Aerodynamics' the word 'Aircraft' and the word 'Wing' might be rejected, even though they are the most frequently occurring words in the document, and the word 'Flutter' might be selected even though, in absolute terms, it is not a high frequency words. Other approaches to automatic indexing use other types of extraction criteria in place of, or along with the statistical criteria, word position in the document, word type, or even the emphasis placed on words in printing—(*e.g.*, boldface and italics)—may all be used as the basis for selection. Subsequently linguistics led the way by pointing out that a number of linguistic processes were essential for the generation of effective content identifiers characterizing natural language texts.

An ideal computerised indexing is one that has the ability to create and modify new subject terms mechanically, by minimising or without the help of human intellectual efforts. As computer can understand only machine code, so it is necessary to translate the information into machine code and in a fixed machine-readable format. Usually, the titles and abstracts are used for the purpose of computerised indexing. However, there are two assumptions:

(*a*) There is a collection of documents; each contains information on one or several subjects.

(*b*) There exists a set of index terms or categories from which one or several of them can describe/represent the subject content of every document in the collection.

Manual Indexing vs. Computerised Indexing

Manual Indexing	*Computerised Indexing*
1. Identifying and selecting keywords from the tile, abstract and full text of the document to represent its content.	1. Key words and/or phrases denoting the subject matter of the document are extracted only from the title and abstract rather than the document's full text.
2. Content analysis of the document is purely a mental process and carried out by the human indexer.	2. The computer does content analysis by following the human instructions in the form of a computer programming.
3. Human indexer makes inferences and judgement in selecting index terms judiciously.	3. Computer cannot think and draw inferences like human indexer and as such, it can select or match keywords, which are provided as input text.
4. Human indexer selects and excludes index terms on the basis of semantic, syntactical as well as contextual considerations.	4. It is possible to instruct a computer through proper programming to select, or exclude a term by following the rules of semantic, syntactical and contextual connotations, like human indexer.

Methods of Computerised Indexing

Keyword Indexing

An indexing system without controlling the vocabulary may be referred as 'Natural Language Indexing' or sometimes as 'Free Text Indexing'. Keyword indexing is also known as Natural Language or Free Text Indexing. 'Keyword' means catch word or significant word or subject denoting word taken mainly from the titles and/or sometimes from

abstract or text of the document for the purpose of indexing. Thus keyword indexing is based on the natural language of the documents to generate index entries and no controlled vocabulary is required for this indexing system. Keyword indexing is not new. It existed in the nineteenth century, when it was referred to as a 'catchword indexing'. Computers began to be used to aid information retrieval system in the 1950s. The Central Intelligence Agency (CIA) of USA is said to be the first organization to use the machine-produced keywords index from Title since 1952. H. P. Luhn and his associates produced and distributed copies of machine produced permuted title indexes in the International Conference of Scientific Information held at Washington in 1958, which he named it as Keyword-In-Context (KWIC) index and reported the method of generation of KWIC index in a paper. American Chemical Society established the value of KWIC after its adoption in 1961 for its publication 'Chemical Titles':

KWIC (Keyword-In-Context) Index

As told earlier, H.P. Luhn is credited for the development of KWIC index. This index was based on the keywords in the title of a paper and was produced with the help of computers. Each entry in KWIC index consists of following three parts:

(*a*) *Keywords:* Significant or subject denoting words which serve as approach terms;

(*b*) *Context:* Keywords selected also specify the particular context of the document (*i.e.*, usually the rest of the terms of the title).

(*c*) *Identification or Location Code:* Code used (usually the serial numbers of the entries in the main part) to provide address of the document where full bibliographic description of the document will be available.

The operational stages of KWIC indexing consist of the following:

(*a*) Mark the significant words or prepare the 'stop list' and keep it in computer. The 'stop list' refers to a list of words, which are considered to have no value for indexing/retrieval. These may include insignificant words like articles (a, an, the), prepositions, conjunctions, pronouns, auxiliary verbs together with such general words as 'aspect', 'different', 'very', etc. Each major search system has defined its own 'stop list';

(*b*) Selection of keywords from the title and/or abstract and/or full text of the document excluding the stop words;

(*c*) KWIC routine serves to rotate the title to make it accessible from each significant term. In view of this, manipulate the

NOTES

title or title like phrase in such a way that each keyword serves as the approach term and comes in the beginning (or in the middle) by rotation followed by rest of the title;

(*d*) Separate the last word and first word of the title by using a symbol say, stroke [/] (sometime an asterisk '*' is used) in an entry. Keywords are usually printed in bold type face;

(*e*) Put the identification/location code at the right end of each entry; and finally

(*f*) Arrange the entries alphabetically by keywords.

Let us take the title 'control of damages of rice by insects' to demonstrate the index entries generated through KWIC principle:

Control of damages of rice by insects	118
Damages of rice by insects/Control of	118
Insects/Control of damages of rice by	118
Rice by insects/Control of damages of	118

In the computer generated index, the keywords can be positioned at centre also.

Variations of KWIC

Two important other versions of keyword index are KWOC and KWAC, which are discussed below:

KWOC (Key-Word Out-Of-Context) Index

The KWOC is a variant of KWIC index. Here, each keyword is taken out and printed separately in the left hand margin with the complete title in its normal order printed to the right.

KWAC (Key-Word Augmented-In-Context) Index

KWAC also stands for 'key-word-and-context'. In many cases, title cannot always represent the thought content of the document co-extensively. KWIC and KWOC could not solve the problem of the retrieval of irrelevant document. In order to solve the problem of false drops, KWAC provides the enrichment of the keywords of the title with additional keywords taken either from the abstract or from the original text of the document and are inserted into the title or added at the end to give further index entries. KWAC is also called enriched KWIC or KWOC. CBAC (Chemical Biological Activities) of BIOSIS uses KWAC index where title is enriched by another title like phrase formulated by the indexer.

Other Versions

A number of varieties of keyword index are noticed in the literature and they differ only in terms of their formats but indexing techniques and principle remain more or less same. They are:

(*i*) **KWWC (Key-Word-With-Context)** Index, where only the part of the title (instead of full title) relevant to the keyword is considered as entry term.

(*ii*) **KEYTALPHA (Key-Term Alphabetical)** Index. It is permuted subject index that lists only keywords assigned to each abstract. Keytalpha index is being used in the 'Oceanic Abstract'.

(*iii*) **WADEX (Word and Author Index).** It is an improved version of KWIC index where the names of authors are also treated as keyword in addition to the significant subject term and thus facilitates to satisfy author approach of the documents also. It is used in 'Applied Mechanics Review'. AKWIC (Author and keyword in context) index is another version of WADEX.

(*iv*) **DKWTC (Double KWIC) Index.** It is another improved version of KWIC index.

(*v*) **KLIC (Key-Letter-In-Context) Index.** This system allows truncation of word (instead of complete word), either at the beginning (*i.e.*, left truncation) or at the end (*i.e.*, right truncation), where a fragment (*i.e.*, key letters) can be specified and the computer will pick up any term containing that fragment. The Chemical Society (London) published a KLIC index as a guide to truncation. The KLIC index indicates which terms any particular word fragment will capture.

Uses of Keyword Index

A number of indexing and abstracting services prepare their subject indexes by using keyword indexing techniques. They are nothing but the variations of keyword indexing apart from those mentioned above. Some notable examples are:

- Chemical Titles;
- BASIC (Biological Abstracts Subject In Context);
- Keyword Index of Chemical Abstracts;
- CBAC (Chemical Biological Activities);
- KWIT (Key Word-In-Title) of Laurence Burkeley Laboratory;
- SWIFT (Selected Words in Full Titles); and
- SAPIR (System of Automatic Processing and Indexing of Reports).

Advantages

1. The principal merit of keyword indexing is the speed with which it can be produced;
2. The production of keyword index does not involve trained indexing staff. What is required is an expressive title coextensive to the specific subject of the document;

NOTES

3. Involves minimum intellectual effort;
4. Vocabulary control need not be used; and
5. Satisfies the current approaches of users.

Disadvantages

1. Most of the terms used in science and technology are standardized, but the situation is different in case of Humanities and Social Sciences. Since no controlled vocabulary is used, keyword indexing appears to be unsatisfactory for the subjects of Humanities and Social Sciences;
2. Related topics are scattered. The efficiency of keyword indexing is invariably the question of reliability of expressive title of document as most such indexes are based on titles. If the title is not representative the system will become ineffective, particularly in Humanities and Social Science subjects;
3. Search of a topic may have to be done under several keywords;
4. Search time is high;
5. Searchers very often lead to high recall and low precision; and
6. Fails to meet the exhaustive approach for a large collection.

Other Methods of Automatic Indexing

Since the KWIC indexing methods various methods generating automatic indexes have been tried. In fact, all attempts at computerized indexing were based on two basic methods: Statistical analysis; Syntactic and Semantic analysis. These are discussed below:

(a) Statistical Analysis

The statistical analysis methods are based on the hypothesis that occurrence of a word in the text indicates its importance. On the basis of this hypothesis a prediction can be made about the subject terms that can be assigned to the document. The computer program can list all the words in a document. The words are grouped by number of occurrences and arranged alphabetically within each frequency. Generally articles, conjunctions, prepositions and pronouns are excluded using a 'stop list' file. Words having same stem can be counted either as the same or as different words. The following methods are adopted in measuring the word significance:

(*i*) **Weighting by location.** A word appearing in the title might be assigned a greater weight than a word appearing in the body of the work.

(*ii*) **Relative frequency weighting.** This is based upon the relation between the number of times the words is used in the document being indexed and the number of times the same word appears in the sample of other documents.

(*iii*) **Use of noun phrase.** Noun and adjective noun phrases can be selected as index terms and these are selected from the title or abstract of the document.

(*iv*) **Use of thesaurus.** A thesaurus can be used to control synonyms and otherwise related terms. In this way, the count of some word types increases as is the separation between 'good' and 'poor' index terms.

(*v*) **Use of association factor.** By means of statistical association and correlation techniques, the degree of term relatedness, that is, the likelihood that two terms will appear in the same document, is computed and used for selecting index terms.

(*vi*) **Maximum-depth indexing.** This procedure indexes a document by all of its content words and weights these words, if desired, by the number of occurrences in the document. In this way, the problem of selecting term is avoided.

(*b*) *Syntactic and Semantic Analysis*

Among the linguistic techniques of interest, the syntactical and semantic analyses are most important in the development of information analysis system needed for computerised indexing. According to Salton, most information analysis systems are based on the recognition of certain key elements, often chosen from a pre-constructed list of acceptable terms, and on the determination of rules by which these basic elements are combined into larger units. The syntactical analysis identifies the role of the word in the sentence, that is, its grammatical class (*i.e.*, parts of speech) and relation among words in the sentence. Whereas semantic analysis helps to establish the paradigmatic or class relations among terms so as to associate words with simple concepts. The main objective of semantic analysis is to identify subject and content bearing words of the document or surrogate text.

Among the linguistic techniques of interest, the following were considered to be of significant:

(*i*) Use of hierarchical term arrangements relating to the content terms in the given subject area can help to expand the standard content description by adding superordinate and/or subordinate terms to a given content description.

(*ii*) Use of synonym dictionaries or thesauri can help to broaden the original context description through a complete class of related terms.

NOTES

(*iii*) Use of syntactical analysis systems capable of specifying syntactic roles of each term and of forming complex content descriptions consisting of term phrases and large syntactic units. A syntactic analysis scheme makes it possible to supply specific content identification.

Use of semantic analysis systems in which the syntactic units are supplemented by semantic roles attach to the entities making up a given content description. Semantic analysis systems utilise various kinds of knowledge extraneous to the documents, often specified by pre-constructed 'semantic graphs' and other related constructs.

Advanced linguistic techniques, that is, the application of computer to analyse the structure and meaning of language led by Noam Chomsky. The linguistic model proposed by Chomsky distinguishes between surface structure and deep structure of a language. By means of transformational grammar, a structure can go through a series of transformations that will exhibit the deep structure. Chomsky found that a purely syntactic transformation could provide a semantic interpretation of the sentence.

Advantages

The advantages of computerized indexing are manifold like level of consistency in indexing can be maintained; index entries can be produced at a lower cost in the long run; indexing time can be reduced; and better retrieval effectiveness can be achieved.

Disadvantages

The main criticism against computerized indexing centres round the fact that a term occurs several times in a document may not always be regarded as a significant term.

File Organisation

Organisation of files in manually operated libraries are expensive, time consuming, labourious and error-prone. Moreover, manual organisation of data/files often leads to duplication of data or data redundancy. Various files maintained in traditional library system are sequential in nature. For example, catalogue cards are arranged in order of search keys (author, title, subject, series etc.) in the form of access points transcribed at the top of the cards.

In computerised indexing system, data elements are stored in suitable digital media and it is possible to manipulate, retrieve and view data elements quite easily. The efficiency of such computerised indexing system largely depends on its file structure.

File organisation forms an important element in computerised indexing. File organization is a technique for physically arranging the records

of a file on secondary storage devices. This is the technique for organisation of the data of a file into records, blocks and access structures.

A file contains data that is required for information processing. These data are about entities. An entity is anything about which data can be stored (*e.g.*, book). The essential properties of an entity are called attributes (*e.g.*, author, title, edition, etc., are attributes of the entity book). Each attribute of an entity is represented in storage by a data item. A data item is assigned a name in order to refer to it in storage, processing and retrieval. Data items are usually grouped together to describe an entity. The data representation in storage of each instance of an entity is commonly called as a record. A collection of related records is called a file.

When data are stored on auxiliary storage devices (*e.g.*, hard disk), the chosen method of file organization will determine how the data can be accessed. The organization of data in a file is influenced by a number of factors, but the most important among them is the time required to access a record and to transfer the data to the primary storage or to write a record or to modify a record.

1. **Sequential File Organisation:** In this technique, records are stored in some predetermined sequence, one after another. One field, referred to as the primary key, usually determines their sequence of order.
2. **Direct File Organisation:** This technique supports direct access (also called random access), in which records can be accessed instantaneously and in any order from the data scattered throughout the disk.
   (*a*) **Relative Addressing:** It is the simplest method of finding a record. Here, a record's primary key is associated with a specific physical storage location and contents of the records are stored in this address.
   (*b*) **Hashing:** It is a method for determining the physical location of a record. Here, the record key is processed mathematically, and another number is computed that represents the location where the record will be stored.
   (*c*) **Indexing:** It is a procedure for locating a record in a file stored randomly throughout the disk. Here, a primary index associates a primary key with the physical location in which a record is stored.
   (*i*) **Ordered Index:** It is based on a sorted ordering of values.
   - **Primary Index:** The records in the indexed file can be stored in some sorted order. If the file containing the records is sequentially ordered, the index whose search

NOTES

key specifies the sequential order of the file is called the primary index.

- **Secondary Index:** Indexes whose search key specifies an order that is different from the sequential order of the file are called secondary indexes.
- **B-Tree Indexes:** It takes the form of a balanced tree in which every path from the root to the tree leaf is of the same length. It eliminates the redundant storage of search key values.

(*ii*) **Hashed Index:** It is based on the values being uniformly distributed using a mathematical function called hash function.

2.7 MACHINE TRANSLATION

Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another. In MT system, the computer program analyses the text in one language-the "source text" and then produces another equivalent text in another language-the "target text"- without human intervention. The translation process, as discussed earlier, involves de-coding the meaning of "source text" and re-encoding the meaning in "target text". However, behind this simple procedure there lies complex cognitive operation. For instance, to de-code the meaning of the text in its entirety, the translator must interpret and analyse all the features of the text, a process that requires in depth knowledge of both, the grammer, semantics, syntax, idioms of the source language as well as the culture of its speakers. The translator needs the same in depth knowledge to re-encode the meaning in the target language. Therein lies the challenge in MT system, how to program the computer to "understand" a text as a human being does and also to "create" a new text in the target language that "sounds" as if it has been written by a human. This problem has been tackled in number of ways. Generally the rule-based methods (such as lexical lookup method, grammer based methods, and meaning based methods), which parse a text, creating an intermediary, symbolic representation, from which the text in target language is generated, have been found successful in machine translation. However, these methods require extensive lexicons with morphologic, syntactic, and semantic information, and large set of rules.

Systems for automatic translations have been under development for over 50 years. The first public demonstration of MT system was held in New York at the head office of IBM in 1954 the system itself, was no more than what today would be called a "toy" system, having just 250 words and translating just 49 selected Russian sentences into

English in the field of chemistry. This demonstration stimulated the financing of MT research not only in U.S. but the worldwide.

Currently the state of machine translation is such that it involves some human intervention, as it requires a pre-editing and post-editing phase. In other words, the MT systems produce output, which must be revised or 'post-edited' by human translators if it is to reach the translation of publishable quality. Sometimes such revisions may be substantial, as MT system produces only a 'draft' translation. However, in fields with limited range of vocabulary and simple sentence structure, machine translations are delivering good results. For example, TAUM, a Canadian MT system, with restricted dictionary of around 20,000 words and expressions, translates weather reports without any human intervention. It is a classic example of an effective and appropriate MT system for a task of limited domain. Since 1977, the system has translated around 15 million words from English to French without any human intervention.

NOTES

Earlier the MT systems were based on 'direct' translations via bilingual dictionaries, with very little analysis of syntactical structures. By 1980s, advances in computational linguistics, allowed much more sophisticated approaches, and a number of systems adopted 'indirect' approach (*e.g.*, 'Interlingua' or 'Transfer') to the task of translation. In these systems, text of source language is analysed into abstract representation of 'meaning' involving successive programs for identifying word structure (morphology), sentence structure (syntax) and for resolving problems of ambiguity (semantics) including component programs to distinguish between homonyms (*e.g.*, English words such as plant which can be botanical or industrial) and to recognise correct semantic relationship. In Transfer approach, there are three basic stages: analysis of the input text into abstract source representation, transfer to a target representation and generation into the output language. In this system three dictionaries are needed: (*i*) A source language dictionary (SD), (*ii*) A target language dictionary (TD), and (*iii*) Transfer dictionary *i.e.*, bilingual dictionary (STD). In transfer approach, the transfer stage requires bilingual component for each language pair, that is, each SL-TL pair. Therefore, in multilingual environment, the number of transfer blocks required, would be equal to the number of languages a MT system covers. The level of transfer differs from system to system. METAL from Siemen Company from Germany is the commercial MT system adopting 'transfer' approach. The other systems adopting the transfer approach are research projects 'Arian' by MT Group GETA in Grenoble and 'Eurotra' funded by European Commission.

Interlingua Approach: In this method the source text is analysed into abstract representation, which is designed to be a kind of language independent 'interlingua' and can serve as an intermediary between

NOTES

large number of natural languages. The translation is in two basic stages: from source language to interlingua and from interlingua into the target language. The example of an experimental interlingua system is the program by Cordier & Mograbhi which translates cooking recipes from French into Arabic.

The most widely known MT systems for mainframe computers are SYSTRAN, METAL, LOGOS and Fujitsu (ATLAS) systems. The SYSTRAN system, originally designed for translation from Russian to English, is now available for 35 language pairs. LOGOS, originally marketed for German to English, is also available for other languages: English into French, German, Italian and Spanish, and German into French and Italian. The Fujitsu ATLAS is for English to Japanese and vice versa. METAL system is for German to English, English to German, German to Spanish, French to Dutch and Dutch to French.

MT systems for personal computers began to appear in the early 1980s. Wielder MicroCat system was the first successful system. Around same time most of the main Japanese computer companies produced systems for translation to and from English, such as the PIVOT system from NEC, the ASTRANSAC system from Toshiba, HICATS from Hitachi, PENSEE from Oki and DUET from Sharp. At the end of 1980s most of the commercial systems on the market presently available appeared. Some of these MT systems for PC are PC-Translator System, Globalink, LogoVista, etc. SYSTRAN, ATLAS, METAL, LOGOS have also brought out PC-based versions. Systran Company offers wide range of PC products such as SYSTRAN Professional, SYSTRAN Personal, SYSTRAN Office Translator, and SYSTRAN WebTranslator. The SYSTRAN MT systems with large dictionary databases and large number of languages, have advantages over other PC based systems.

The demand for translations of electronic texts on the Internet, such as web pages, electronic mail and even electronic 'chat' lists, is developing rapidly. Many MT vendors like Systran, Logos, Globalink, Fujitsu, JICST and NEC have been providing network –based translation services for on-demand translation with or without human editing. Many new companies have also launched their products specifically for Internet. LANT in Belgium has launched its multilingual service for translation of electronic mail, web pages and attached files. MTSU in Singapore is providing large-scale translation over the Internet for many customers worldwide. Many Japanese companies have been providing MT software products for translating web pages. Even Internet services are adding translation facilities. For example, AltaVista is offering translation facility, for translating French, German, and Spanish into and from English, on the Internet. Equally significant has been the use of MT for electronic mail and for 'chat rooms'. In 1994 the CompuServe service introduced automatic translation from to English

NOTES

and French, German or Spanish for messages on one of its forums. It became so popular that within next two years the facility was extended to two other online services. Now thousands of messages are being translated. CompuServe has introduced its own translation service on the Internet for longer documents either as unedited 'raw' MT or with optional human editing. CompuServe also offers MT as standard for all its e-mail. Many of the MT systems offer free translation facility on the web. Listed below are some of these sites. (http://www.compuserve.com/)

http://www.google.com/language_tools (uses SYSTRAN software)

http://www.freetranslation.com/

Another area of MT research is the development of systems for spoken language translation *e.g.*, in telephone conversation and in business negotiations. In Japan a joint government and industry company ATR, was established in 1986 near Osaka for research on speech translation. It is now one of the main centres for automatic speech translation. The aim is to develop a speaker- independent real-time telephone translation system for Japanese to English and vice versa, initially for hotel reservation and conference registration transactions. Other speech translation projects are JANUS system in Carnegie-Mellon University and Karlsruhe in Germany. The researchers are collaborating with ATR in a consortium (C-STAR), each developing speech recognition and synthesis modules for their own languages (English, German, Japanese).

Tools for Translators

It has been observed that professional translators spend considerable amount of their time consulting technical (monolingual, bilingual and multilingual) dictionaries, to find suitable technical terms while translating S&T documents. They have long been wanting to have sophisticated computer-based translation tools to speed up translation process. The development of translation tools became feasible, firstly with the availability of real time interactive computer environments in late 1960s, then with the appearance of word processing in the 1970s and of microcomputers in 1980s and, subsequently, with intra organisational networking and the development of large computer storage capacities. Since the early 1990s the translators could have these translation tools in the form of translation workstations. The translation workstations offer translators the opportunities of making their work more productive without taking away intellectual challenge of translation. Translation workstations combine access to dictionaries and terminological databanks, multilingual word processing, the management of glossaries and terminology resources, appropriate facilities for input and output of texts (*e.g.*, OCR scanners, electronic transmission, high-class printing) and translation memories.

Translation Memory (TM) is a software programe designed as an aid for human translators. A translation memory consists of a database of text segments in source language and their translation in one or more languages. These segments may be individual words or multiword phrases. Using TM a translator can translate save and reuse translated sentences and passages. When translator comes across similar or identical material, he can reuse the previously translated material. Even if there is not exact match, the version displayed may be used with minor changes. These systems are very suitable to help translators with technical documentations and documents containing specialised vocabularies. Among other advantages of TM are:

(*i*) Consistency in common definitions, phrases and terminology when a number of translators work on a single translation project;

(*ii*) Speeding up overall translation process; and

(*iii*) Making the translation process cost effective for long term translation.

There are 4 main vendors of translation workstations: (*i*) Trados, (*ii*) STARAG Company (Transit), (*iii*) IBM (The Translation Manager), and (*iv*) LANT in Belgium (The Eurolang Optimiser previously sold by SITE in France).

Translators Associations

Indian Scientific Translators Association (ISTA) was established in 1962. Dedicated to the cause of promotion of scientific translation in India, the association has been making persistent efforts to bring S&T translation to the focus of public and government attention through talks, discussions, seminars, surveys, workshops, publications, etc. The main objectives of ISTA are to:

- promote facilities for scientific translation in India;
- endeavour to improve the status and service conditions of scientific translators;
- promote training facilities for scientific translation and take such measures as would lead to the maintenance of a high standard of scientific translation;
- convene conferences or conduct seminar on scientific translation;
- cooperate with national and international organisations with similar objectives;
- bring out publications which will tend to the realisation of the objectives of ISTA; and
- do all such activities as are incidental or conducive to the attainment of the objectives of ISTA.

Since 1972, the association has been publishing JISTA, Journal of Indian Scientific Translators Association. Published annually, JISTA covers articles and news items on theoretical and practical aspects of translation, training of translators, linguistic issues, scientific terminology problems, machine translation, updates on science and technology, etc. The association celebrates 'International Translation Day' on 30 September every year (since 1992). The focus has been on French translation, Korean translation, machine translation and machine translation tools, etc. The association has brought out 'Directory of S&T Translators in India' and maintains it on its website (http://education.vsnl.com/ista/).

In brief, in this section we have discussed the need for translation services, how to organise translation service in a documentation or information centre, and current scenario of translation facilities in India in the field of science and technology, humanities, and social sciences. A brief account of bibliographical control of translations at international level is provided. Described in detail, research in machine translation; major MT systems for mainframe, personal computers, and for Internet; and computer based tools for translators such as translation workstations, translation memories, etc. A brief account of Indian Scientific Translators Association is also provided. Website addresses of some of the websites offering free-to-use machine translation facility, on the web are listed.

2.8 THESAURUS

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

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

NOTES

Definition

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

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

Functions

The functions of a thesaurus are:

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

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

Thesauri have been designed in specific fields/areas by international organisations other than libraries. Some examples include: Thesaurus of Engineering and Scientific Terms (TEST), ERIC Thesaurus for Education, Thesaurus of American Psychological Association (APA), Unesco Thesaurus, etc.

Construction

Thesaurus construction is a very specialized activity. Anyone involved in its construction should have a sound knowledge of the subject and

should be logical and have organisational capabilities. The steps for construction of a thesaurus are as follows [*Lancaster, 1985*]:

NOTES

(a) Need Analysis

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

(b) Gathering of Terms

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

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

(c) Organisation of Terms

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

(d) Organisation into Hierarchies

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

(e) Creation of Alphabetical Thesaurus

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

NOTES

(f) Presentation of Thesaurus

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

(g) Evaluation

Once the thesaurus is compiled it needs to be evaluated to assess its retrieval effectiveness.

(h) Maintenance

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

(i) Use of Computers

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

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

2.9 SUMMARY

- An indexing language is an artificial language consisting of a set of terms and devices for handling the relationship between them for providing index description.
- Theories also supply a rationale for, or an argument against, current practices in subject indexing.
- The criteria at best enable greater agreement between indexers about concepts that should be indexed.
- An indexing language is considered to be of high specificity if minute concepts are represented precisely by it.
- The Universal Decimal Classification (UDC) is a worldwide popular general classification scheme covering all fields of knowledge.

- The UDC can be applied in different contexts that make it more flexible than other general classification schemes.
- The main tables of UDC comprise of ten main classes that represent whole of universe of recorded knowledge.
- The common auxiliary tables provide notation for relationships or recurring concepts.
- The filing order of UDC symbols is based on a progression from the general to the specific.
- The citation order is governed by the rules and conventions of any classification scheme for such purposes as displaying the concepts in the UDC schedules.
- The keyword indexing is based on the natural language of the documents to generate index entries and no controlled vocabulary is required for this indexing system.
- Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another.

2.10 REVIEW QUESTIONS

1. Discuss the processes of subject indexing.
2. What do you mean by 'Exhaustivity' and 'Specificity' in indexing?
3. What is machine translation? Describe different types of approaches used for machine translation.
4. Enumerate the structural differences between a Thesaurus and a List of Subject Headings.

2.11 FURTHER READINGS

1. Chan, Lois Mai (1994). Cataloguing and classification: an introduction. 2nd edition. New York: McGraw-Hill.
2. Foskett, A.C. (1996). The subject approach to information. 5th ed. London: Library Association Publishing.
3. Gakhar, A.P. (1982). Librarian's guide to Broad System of Ordering (BSO). New Delhi: Metropolitan.
4. Hunter, Eric J. (2002). Classification made simple. 2nd edition. Aldershot: Ashgate.
5. Krishan Kumar (1988). Theory of library classification. 4th edition. New Delhi: Vikas.
6. Marcella, Rita and Newton, Robert. (1994). A new manual of classification Aldershot: Gower.

NOTES

UNIT III INFORMATION STORAGE AND RETRIEVAL (ISAR) SYSTEMS

★ STRUCTURE ★

LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain information retrieval systems
- know about information retrieval features of OPACs
- describe the file organisation in ISAR systems
- know about users and their information needs
- explain objectives of ISAR systems.

3.1 INTRODUCTION

The term information retrieval was coined by Kelvin Mooers over 50 years ago, but it gained popularity in the research community after about a decade, in the early sixties, when computers were being introduced in information handling. The term information retrieval was then being used to mean retrieval of bibliographic information from stored document databases. Truly speaking, these information retrieval systems were document retrieval systems; they were designed to retrieve information about the existence (or non-existence) of bibliographic documents relevant to a user's query. In other words, early information retrieval systems were designed to retrieve an entire document (a book, an article, etc.) in response to a search request. While this is very much what today's information retrieval systems do, over the

NOTES

years many advanced techniques have been developed and applied to design information retrieval systems. These techniques and other modules on information retrieval will be discussed in this Unit. Over the years the connotation of information retrieval has changed and it has been variously termed by information professionals and researchers, some of which include: information storage and retrieval, information organization and retrieval, information processing and retrieval, text retrieval, information representation and retrieval, and information access.

3.2 INFORMATION RETRIEVAL SYSTEMS

An information retrieval system is designed to analyse, process and store sources of information and retrieve those that match a particular user's requirements [*Chowdhury, 2004*]. Modern information retrieval systems can either retrieve bibliographic items, or the exact text that matches a user's search criteria from a stored database of full texts of documents. Although information retrieval systems originally meant text retrieval systems since they were dealing with textual documents, modern information retrieval systems deal not only with textual information but also with multimedia information comprising text, audio, images and video. While many features of conventional text retrieval systems are equally applicable to multimedia information retrieval, the specific nature of audio, image and video information have called for the development of many new tools and techniques for information retrieval. Thus, modern information retrieval systems deal with storage, organization and access to text, as well as multimedia information resources.

The concept of information retrieval pre-supposes that there are some documents or records containing information that have been organized in an order suitable for easy retrieval. The documents or records that we are concerned with contain bibliographic information which are quite different from other kinds of information or data. We may take a simple example. If we have a database of information pertaining to an office, or a company, all we have are the different kinds of records and related facts, like names of employees, their positions, salary, and so on, or in the case of a manufacturing company, names of different items, prices, quantity, and so on. The retrieval system here is designed to search for and retrieve specific facts or data, like the salary of a particular manager, or the price of a perfume, and so on. The major objective of an information retrieval system, on the other hand, is to retrieve the information – either the actual information or through the documents containing the information surrogates – that fully or partially match the user's query. Thus, the search output may contain bibliographic details of the documents that matches the query, or the

NOTES

actual text, image, video, etc. that contain the required information. The database in case of an information retrieval system may contain abstracts or full texts of documents, like newspaper articles, handbooks, dictionaries, encyclopedias, legal documents, statistics, etc., as well as audio, images, and video information.

Whatever may be the nature of the database – bibliographic, full-text or multimedia – the system pre-supposes that there is a group of users for whom the system is designed. Users are considered to have certain queries or information needs, and when they put forward their requirement to the system, the latter should be able to provide the necessary bibliographic references of those documents containing either the required information, or the actual text in the case of a full-text retrieval system. Alternative models of (knowledge-based) information retrieval seek to provide the user with the information directly rather than just the citations, the abstract or the full text.

Databases

An information retrieval system deals with databases, and so does a database management system. So, what is the difference between an information retrieval system and a database management system? Before we discuss these differences, we need to have some basic idea of a database and its various components, types, etc. which are discussed in the following sections.

Data

The data is discrete fact, when processed it becomes information. However, in the context of Information Retrieval System, we may consider information as a logical set of data. The word 'data' refers to a set of given facts. Information in a form that can be processed by a computer is called data. The term data has for long been used to refer to scientific measurements, but words also constitute data just as numbers do. A list of names is data, a set of keywords is data, a doctor's record of his patients is data, and figures relating to temperature, humidity, etc., or sales of a company, are data.

The Database

A database can be conceived as a system whose base, whose key concept, is simply a particular way of handling data. In other words, a database is nothing more than a computer-based record-keeping system. The overall objective of a database is to record and maintain information. The Macmillan Dictionary of Information Technology [Longley and Shain, 1989] defines a database as 'a collection of interrelated data stored so that it may be accessed by authorised users with simple user-friendly dialogues'. The Chambers Science and Technology Dictionary [*Walker, 1988*] provides a more simple definition of a database: 'a collection of structured data independent of any particular application'.

NOTES

It may be noted from the above definitions that a database contains some data that are structured and integrated. Ellingen [1991] defines a database as 'a collection of information that can be searched as a single entity'. According to Oxborrow [1989], a database can be considered as 'an organised collection of related sets of data, managed in such a way as to enable the user or application program to view the complete collection, or a logical subset of the collection, as a single unit'.

From the above definitions we can simplify the definition of a database as an organised collection of related sets of data that can be accessed by more than one user by simple means and can be searched to reveal those that touch upon a particular need [*Chowdhury, 2004*]. In the computer world we frequently deal with files, which are the outer identifying boundary or a sort of folder containing data. Thus, a file is equivalent to an ordinary address book, if we are talking about names and addresses. A file in a computer is given a unique name by which it is addressed.

Records and Fields

A record is a collection of related information. A database is an organised collection of units of information, and each unit of information in a database is called a record. A record is generally what a user wants to find out while searching a database. An example of a record is the main entry in a library's catalogue, which describes the book's title, author, subject, etc. A collection of database records constitutes a database file. Identifying what the record is to be is one of the early tasks in designing a database. If the database is a bibliographic one, the bibliographic information about each document is the unit of information or record.

A stored record is a named collection of associated stored fields [Date, 1981]. Each record is made up of particular segments or elements of information, each of which is called a field. A field holds a particular type of information within a record that can be separately addressed. The different items of information in a bibliographic record may be author, title, subject heading, etc. Thus, the different fields in a bibliographic record can be the 'author field' containing name(s) of author(s), the 'title field' containing the title, and so on. A field may be subdivided into still smaller units called subfields. For example, if 'imprint' of application is regarded as a field in a bibliographic database, then the different components of the imprint – the publisher's name, place of publication, and date of publication – can be called subfields.

A record is, thus, composed of fields and subfields. Identifying what fields and subfields are to be included in each record is an important task in the database design process. Each field is given a unique identifier, at the design stage, called field tag, which is then used throughout for data input, editing, searching, printing, and so on.

Several standards have been developed to help the designers of information retrieval systems in this regard. For example, in case of an online catalogue, or more specifically OPAC (Online Public Access Catalogue), as they are called, standard bibliographic format like MARC21 [2002] (MARC stands for Machine Readable Catalogue or Cataloguing; several different types of MARC formats have been developed and MARC21 is the most recent and the most heavily used MARC format), CCF (Common Communication Format) [1992], and so on, specify the fields and the corresponding field tags to be used while preparing catalogue entries for bibliographic items.

Properties of Databases

A database is designed to avoid duplication of data as well as to permit retrieval of information to satisfy a wide variety of user information needs. Major properties of a database can be summarised as follows:

- it is integrated with provisions for different applications;
- it eliminates or reduces data duplication;
- it enhances data independence by permitting application programs to be insensitive to changes in the database;
- it permits shared access;
- it permits finer granularity; and
- it provides facilities for centralised control of accessing and security control functions.

Kinds of Databases

In discussing databases, it is sometimes useful to classify them by the type of data record contained and sometimes by subject coverage. The two major divisions are reference databases and source databases. Reference databases lead the users to the source of the information: a document, a person or an organisation. They can be divided into three categories:

(*a*) bibliographic databases, which include citations or bibliographic references, and sometimes abstracts of literature;

(*b*) catalogue databases, which show the catalogue of a given library or a group of libraries in a network; and

(*c*) referral databases, which offer references to information such as the name, address, specialisation, etc., of persons, institutions, information systems, etc.

Source databases provide the answer with no need for the user to refer elsewhere. These databases contain the information sought for in electronic form and, therefore, the user can get access to the information instantly as a result of a search. Source databases can be grouped according to their content, for example,

(*a*) numeric databases, which contain numerical data of various kinds, including statistics and survey data.

(*b*) full-text databases, which contain the full text of documents.

(*c*) text-numeric databases, which contain a combination of textual and numerical data, such as a company annual report and handbook data.

Bibliographic databases form the basis of most of the information retrieval systems available today, be they home-grown or available on CD-ROM or through online access. Bibliographic databases can be divided into five broad categories:

(*a*) large discipline-oriented databases;

(*b*) interdisciplinary databases with coverage based on key or core journals;

(*c*) cross-disciplinary databases;

(*d*) smaller, more specialized databases serving a particular technology or application area; and

(*e*) databases covering specific types of publication.

However, there could be many more kinds of bibliographic databases, such as:

- *Specific subjects/disciplines:* CASearch, BIOSIS, ERIC, MEDLINE, ENERGYLINE, LISA, ISA, and so on;
- *Multidisciplinary:* SCI SEARCH, SOCIAL SCISEARCH;
- *Mission-oriented:* NASA;
- *Problem-oriented:* ENVIROLINE, TOXLINE;
- *Referral:* Foundations Directory, Fine Chemicals Directory, Ulrich's International Periodicals Directory;
- *Factual:* PTS Forecasts, CARIS/FAO (Ongoing Research);
- *Textual references:* DRUGLINE; and so on.

Many of these databases are available online, accessible through the web, and CD-ROM versions.

Information Retrieval vs. Database Management Systems

The technology that helps to process and manipulate data of various kinds is broadly termed as database management technology, and the resulting software packages are known as Database Management Systems (DBMSs). A database management system stores and retrieves discrete data elements that are structured, as opposed to a typical information retrieval system that is designed to deal with unstructured data *e.g.*, the full texts of documents.

Typically a search in a database management environment produces one or more records that are stored in the database. One may argue that an information retrieval system also stores discrete data elements,

like author, title, keyword, etc., in the form of a structured database. While this is true, an information retrieval system also handles unstructured data, for example a large chunk of text, and this is where a typical database management system differs from an information retrieval system. Many more differences between the two systems can be noticed especially in the search and retrieval aspects. For example, in a typical database management search, we expect to retrieve discrete data, *e.g.*, the price of an item, date of birth of an employee, and so on, whereas in information retrieval search we retrieve an entire document or part of it containing the information required by the user. The major differences between a typical database management system and an information retrieval system are shown in Table 3.1

Table 3.1: Difference between Information Retrieval Systems and Database Management Systems

Information Retrieval Systems	*Database Management Systems*
Designed to deal with unstructured data	Deals with structured data
An item may be retrieved if it exactly or partially exactly matches a query	An item will be retrieved only when it matches the query
Queries are usually language-based, e.g., a keyword, an author name etc.	Queries are mostly value-based, *e.g.*, salary or date of birth of a person
Vocabulary is very important and usually some vocabulary control tools are used	No vocabulary control tool is required
A number of advanced search techniques are used, for example proximity search	Exact match of search term and field value is expected

Information Retrieval Systems: Purpose, Components and Functions

An information retrieval system is designed to retrieve the documents or information required by the user community. It should make the right information available to the right user at the right time. Thus, an information retrieval system aims at collecting and organizing information in one or more subject areas in order to provide it to the user as soon as asked for. Belkin [1980] presents the following situation which clearly reflects the purpose of information retrieval systems:

(*a*) a writer presents a set of ideas in a document using a set of concepts;

(*b*) somewhere there will be some users who require the ideas but may not be able to identify those. In other words, there will be some persons who lack the ideas put forth by the author in his/her work; and

(*c*) information retrieval systems serve to match the writer's ideas expressed in the document with the users' requirements or demands for those.

Thus, an information retrieval system serves as a bridge between the world of creators or generators of information and the users of that information. The information resources are processed, indexed and stored in an appropriate way. The users interact with the system through a user interface. The user queries, submitted through the interface are matched with the index and the matching items are retrieved. A number of activities are involved in the processes of information processing, indexing and matching.

The retrieval process begins with a user query. A user with an information need, interacts with the information retrieval system through the user interface and submits a query. A search query may contain a simple keyword or a phrase, or may contain more than one keyword or phrase combined with some search operators. The retrieval system matches each search term with the inverted index file, and retrieves the matching items. Although this is the basic process in an information retrieval operation, the specifics of each of these activities may be quite complex and depend on the retrieval system, or the retrieval engine, as they are now called. A number of tools may also be used in one or more stages. For example, vocabulary control tools like thesauri, and/or machine translation tools, may be used in the indexing and/or retrieval process.

Indexing and Information Representation

In a typical information retrieval environment, the users queries are not matched with the documents per se; instead, they are matched with an index file. The actual documents are stored in a separate sequence, and once a match is found between an index term and a user search term, the pointer from the index file is followed to retrieve the document.

The elementary units of a text retrieval system are the document records. Each document record comprises of a number of fields and subfields, each one of which contains a particular unit of information – author's name, publisher's name, title, keyword(s), class number, ISBN, and so on. The document record may also contain an abstract or full text of the document concerned. A text retrieval system is designed to provide fast access to the records through any of the sought keys or access points. This means that there should be a mechanism for fast access to the document records. What should the basic mechanism be for accessing the document records through some key values – by

NOTES

chosen keyword(s), or by author's name, say? To answer this question, we should first understand how document records are physically stored in the computer.

Document records are stored one after another in the computer memory: this is actually the virtual structure of the database file. Imagine a text database that stores a few, say ten, document records. Now, suppose a user wants to check if there is a document in the database that is written by G.G. Chowdhury; another user wants to know if there is a book on Internet. What would the user's approach be? The simplest way would be to open each document record one after another and to check each and every field; if there is a match then it retrieves that document. This process continues until all the document records have been checked. It may be a very simple approach, but one can very well imagine that this will be an extremely slow process even for a faster computer when the text database is relatively large, and will be an impossible proposition for a database that has some hundreds or thousands of document records or more.

What is the solution then? How can we retrieve the desired document records? Let's take a common example. What do we do when we want to locate a particular term or phrase, say the word 'computer', or the phrase 'information retrieval', in a book? Do we start from the first line in the first page and continue up to the last line in the last page of the book? No: we use a simple tool – the back-of-the-book index. What is such an index? It is a simple alphabetical list of all the potential index terms, drawn from the text of the book, each having a pointer showing the occurrence(s) of the terms. Thus, we look into the index file with the required search term, locate it and then move to the page(s) indicated for the actual information. A similar approach is taken in a text retrieval system: an index file is created that contains all the potential index terms arranged in an appropriate order. This index file is called an inverted file. Users looking for particular information are required to consult the inverted file first, which then leads to the main database where the document records are stored.

Like inverted file, two more file structures also exist for representation and access to information. These are sequential file and indexed sequential file.

Sequential File: In a sequential file the records are arranged in order of a key field and the computer can use a searching technique, like a binary search, to access a specific record. A sequential file is designed for efficient processing of records in sorted order on some search key. In this file structure, records are chained together by pointers to permit fast retrieval in search key order. Pointer points to next record in order. Records are stored physically in search key order (or as close to this as possible). This minimises number of block accesses.

Indexed Sequential File: Indexed sequential file is a type of file access in which an index is used to obtain the address of the block containing the required record. In indexed sequential files each record of a file has a key field which uniquely identifies that record. It has an index consists of keys and addresses. Indexed sequential files are important for applications where data needs to be accessed either sequentially or randomly using the index.

Example: A library may store details about its users as an indexed sequential file. Sometimes the file is accessed sequentially: when the whole of the file is processed to produce overdue statistics at the end of the month.

Randomly: may be a user changes address, or a lady user gets married and changes her surname. An indexed sequential file can only be stored on a random access device, *e.g.*, magnetic disc, compact disc (CD).

Inverted File

In an inverted file system of text retrieval, each database consists of two files. One is the text file, which contains what we would expect to find, that is the document records in their normal form – the form in which they are entered into the database. The other is the inverted file, which contains all the index terms, drawn automatically from the document records according to the indexing technique adopted for the purpose. Each index term in the inverted file is associated with a pointer which shows the record number in which the index term occurs.

The indexing technique, *i.e.*, the technique adopted to draw index terms from the records, determines the order in which index terms will appear in the index file. Different techniques may be required for the purpose: for example, index entries may be required for:

- each and every term occurring in a given field, for example, all the words occurring in the title field. However, there is a risk; some unwanted terms, like 'a', 'an', 'and', 'the', etc., occurring in the document titles may also be indexed. To avoid this problem, text retrieval systems usually incorporate a stop-word file which prevents unwanted terms from being indexed
- the whole field as it is, for example, the full title as it occurs in the document record
- each occurrence of a repeatable field, for example, names of all the authors
- some selected words or phrases from a field or subfield, for example, *some terms and phrases occurring in the title field, etc.*

Thus, for each significant index term in the database the inverted file contains an entry along with a reference list which specifies position(s) in the database where that term appears. Therefore, in an inverted

NOTES

file system, the searcher first consults the index file, which then refers to the position in the main text database where the desired record appears. The inverted file system is, thus, an example of indirect file access. If the terms are arranged alphabetically in the inverted file, then the file represents an example of indirect sequential file organization. An inverted file may contain a lot of other information along with each entry, such as the number of occurrences of the term in a given record or position information, such as the field in which the term/phrase occurs, where the term/phrase occurs in a given sentence/paragraph, and so on. Index entries are drawn from all four sample document records for the author, title, publisher, and keyword fields. Titles have been indexed as they are, while each occurrence of the author and the keyword field in the document records has been indexed.

Document records

Document no: 1
Author: Cunningham, M.
Title: File structure and design
Publisher: Chartwell-Bratt
Year: 1985
Keywords: File structure; File organization

Document no: 2
Author: Tharp, A.
Title: File organization and processing
Publisher: John Wiley
Year: 1988
Keywords: File structure; File organization

Document no: 3
Author: Ford, N.
Title: Expert systems and artificial intelligence
Publisher: Library Association
Year: 1991
Keywords: Expert systems; Artificial intelligence; Knowledge-based systems

Document no: 4
Author: Charniak, E.; McDermott, D.
Title: Introduction to artificial intelligence
Publisher: Addison-Wesley
Year: 1985

Fig. 3.1. *Sample document records*

NOTES

Index file

4 40 1 1	Addison-Wesley
3 60 1 2	Artificial Intelligence
4 60 1 1	Artificial Intelligence
4 20 1 1	Charniak, E.
1 40 1 1	Chartwell-Bratt
1 20 1 1	Cunningham, M.
3 60 1 1	Expert Systems
4 60 1 2	Expert Systems
3 30 1 1	Expert Systems and Artificial Intelligence
1 60 1 2	File Organization
2 60 1 2	File Organization
2 30 1 1	File Organization and Processing
1 60 1 1	File Structure
2 60 1 1	File Structure
1 30 1 1	File Structure and Design
4 30 1 1	Introduction to Artificial Intelligence
3 60 1 3	Knowledge-based Systems
3 40 1 1	Library Association
4 20 1 2	Mcdermott, D.
2 20 1 1	Tharp, A.

Fig. 3.2. *Sample inverted index file*

The field tag is used to denote the field where the given term/phrase occurs. This information is used in field-specific searches. Similarly, the position information is used for proximity or adjacency searching. Other types of information may also be stored along with each entry, and each such item of information facilitates a particular type of search. Nevertheless, the more such information is added to each entry, the more bulky the inverted file becomes, and therefore takes more storage space and processing time. In this example, a user looking for a term 'expert systems' will retrieve two records, document numbers 3 and 4 from the database, while another user looking for a book written by 'Tharp, A.' will retrieve book number 2. A complex query with search terms combined by Boolean operators will follow the same path. For example, a user with a query 'expert systems OR file organization' will retrieve all four document records, while the query 'artificial intelligence AND knowledge-based systems' will retrieve document record number 3. In the first example, as the search terms are joined by the logical operator 'OR', the system will consult the inverted file for each term and then will merge the document numbers retrieved in each case, while in the second case, because the terms

are joined by the logical operator 'AND', the retrieved document numbers for both terms will be matched to locate the common document numbers, that is the ones where both terms are present. Figure 3.1 shows that an index term may occur in several document records, and in each case, several items of information, such as its frequency of occurrence, field(s) in which it has occurred, position information, and so on, have to be stored in the index file. Thus, conceptually the structure of an inverted file may look like the one shown in Figure 3.2.

Access to Inverted Files

The user may pose a single key query or a multiple key query. In the former case, the value of a single search key (say the name of the author) is used as the retrieval criterion, whereas in a multiple key search a number of search keys (say the name of the author, subject name, date of publication, and so on, as in the query 'papers written by Salton on information retrieval systems between 1980 and 1990'). For single key searches, the whole file can be maintained in an order according to the value of the given single set of keys. In a telephone directory, for example, users search through the names of subscribers and therefore the names of subscribers are arranged in alphabetical order. File access in multi-key searches is complicated by the fact that it is not possible to order the file simultaneously in accordance with the values of the different search keys. For example, a users' file in a library can be arranged according to the name of the user, occupation or specialisation, address or department, and so on, and in each case the resulting arrangement of the records within one field will be different from the other.

In the case of a multi-key search, a principal key is to be identified and the file can be ordered in accordance with the values of that key. When the principal key is used as part of a search statement, the subsection of the file corresponding to the given principal key value can then be isolated and subjected to a separate search based on the values of any secondary keys also included in the search query.

A catalogue of a library can be considered as a multi-key file, where the keys are the author, title, publisher, subject, etc. In such a file, the principal key is usually the author, *i.e.*, the file is ordered in accordance with the name (surname) of the authors. From each record in the main file there may be a number of pointers giving access to secondary keys, like publisher, title, etc. A simple file of authors and publishers can be ordered according to the author's name as the principal key, with a sparse index giving access to a chain of pointers for each publisher name. Documents published by a given publisher can be found by following the pointer chain. Pointer chains can be provided for all secondary keys in addition to the primary keys attached to the records; each given record can be traced through the pointer chain for any of the keys. This type of record organisation is known as a multi-list.

Multi-list organisation, however, becomes too time-consuming when each query key is attached to a large number of records. One solution to this might be to use large indexes that provide one pointer for each record exhibiting a given key value. Such an index is called an inverted index or an inverted file. Inverted files are widely used in operational information retrieval situations. The advantage of using inverted files is that such files allow extremely rapid search and retrieval operations, based only on the information provided in the index rather than on data from the main record file.

One important issue for the inverted file system is the size of the index file. If each and every term occurring in the document database is indexed, then size of the index file will be quite large, equal to that of the main document database. Therefore, in order to facilitate fast searching, we need to have a method that allows fast access to the terms/phrases in the inverted file. In other words, we need to have an efficient file organization technique.

Vocabulary Control

Vocabulary control is one of the most important components of an information retrieval system. An information retrieval system tries to match user queries with the stored documents (the inverted index file to be precise) and retrieves those that match. In order to match the contents of the user requirements (the search terms) with the contents of the stored documents (the index entries), one must follow a vocabulary that is common to both. In other words, user requirements need to be translated and put to the retrieval systems in the same language (using the same terms, for example) as was used to express the contents of the document records. This leads us to the concept of using a standard or controlled vocabulary in an information retrieval environment.

Indexing may be thought of as a process of labelling items for future reference. Considerable order can be introduced into the process by standardizing the terms that are to be used as labels. This standardization is known as vocabulary control, the systematic selection of preferred terms.

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

NOTES

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

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

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

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

A number of vocabulary control tools have been designed over the years. They differ in their structure and design features, but they all have the same purpose in an information retrieval environment. A number of software packages are now available that allow the record creator to automatically switch to one or more chosen online vocabulary control tools in order to select appropriate terms for representing the document in hand. For example, OCLC's Connexion (an integrated cataloguing suite of tools) and OCLC's CatExpress (simple copy cataloguing suite of tools) provide such facility. This helps in a number of ways – the document records do not only contain a number of terms that are representative of the contents of the document, but these are also standardized (in terms of their usage, spelling, form, and so on) and are likely to be chosen by the user for searching purposes. Similarly, there are programs available by which end-users may go to the appropriate page of a particular online vocabulary control tool in order to choose the most appropriate term(s) for preparing the search expression. Vocabulary control tools also help end-users modify their previously formulated search expressions by either widening or narrowing down the search expressions.

Vocabulary Control Tools

As the name suggests, these are the tools used to control the vocabulary of indexing and retrieval. What an indexer and an index user need is a set of guidelines for the proper selection of terms. Syntactic structures are devices that provide these guidelines by showing the relationships among terms or concepts, and they fall into two major categories: (*i*) classification schemes, and (*ii*) subject heading

lists and thesauri. A combination of the two categories has also been developed.

Classification schemes, being tools for organising knowledge, could be of great help for vocabulary control but the main body of classification schemes is organised in an artificial language (called notations which may contain numbers, alphabets, punctuation marks, or a combination of them) whereas for vocabulary control we need natural language representation. Indexes to classification schemes could serve the role of vocabulary control but here terms appear alphabetically and thus the logical (semantic) organisation of knowledge is not available. Some attempts have been made to combine the features of the main arrangement in classification schemes with those that appear in the index to the classification scheme to generate some kind of faceted or classified thesaurus such as thesauro-facet.

Subject heading lists were initially developed to prepare entries/headings in a subject catalogue that could replicate the classified arrangement of document records. Therefore, they include rather broader subject terms or headings. On the other hand, thesauri have been developed on specific subject fields with a view to bringing together the various representations of terms (synonyms, spelling variants, homonyms, etc.) along with an indication of a mapping of that term in the universe of knowledge by indicating the broader (superordinate), narrower (subordinate), and related (coordinate and collateral) terms. However, this distinction has gradually faded and the latest Library of Congress subject headings list indicates the terms' features as shown in normal thesauri.

Controlled vs. Natural Language Indexing

Controlled indexing languages are those in which both the terms that are used to represent subjects and the process whereby terms are assigned to particular documents are controlled or executed by a person. Normally there is a list of terms—a subject headings list or a thesaurus, that acts as the authority list in identifying terms that may be assigned to documents, and indexing involves the assignation of terms from this list to specific documents. The searcher is expected to consult the same controlled list during formulation of a search strategy. In natural language indexing, any term that appears in the title, abstract or text of a document record may be an index term. There is no mechanism to control the use of terms for such indexing. Similarly, the searcher is not expected to use any controlled list of terms.

Whether to use a controlled vocabulary or to use natural language indexing has been an age-old debate in information retrieval.

NOTES

Table 3.2: The Four Eras of Debate on Controlled Vs. Natural Language Indexing

Era One –	controlled vocabulary
Era Two –	comparisons of natural and controlled language: major experimental studies noted that natural language can perform as well as controlled vocabulary, but other factors, such as the number of access points, are also significant.
Era Three –	many case studies of limited generalizability. Searching online databases was considered. It was noted that the best performance can be achieved by a combination of controlled and natural language; the number of access points was reaffirmed to have a significant effect; full-text and bibliographic databases were noted to have produced different results.
Era Four –	new advances in user-based systems including OPACs. The value of controlled vocabulary in the context of user-friendly interfaces and the development of knowledge bases were noted.

Aitchison and Gilchrist [2000] provide a detailed comparison of natural and controlled language indexing. However, despite much debate extending over more than a century, together with a range of research projects, information scientists have failed to resolve the issue concerning the relative merits and demerits of controlled and natural language. Evidences produced by practice and tested research suggest that controlled language and natural language may be used in conjunction with one another.

Searching

While searching for information in a database, users may approach with some keys. For a bibliographic database, such keys can be author name, title, ISBN, subject keywords, etc. In a non-bibliographic database, these keys will depend on the nature of the database concerned.

In a bibliographic information retrieval environment, searches can be divided into two main classes: known item search and unknown item search. A known item search is the one where the user knows something about the item being sought. This may be any key like author, title, publisher, ISBN, and so on. In such a case, user can enter the appropriate key and can get the full details of the item concerned. For example, the user can enter the author name to retrieve the full record. However, very few users actually know about the author, title, etc., of the item that he/she might need at a given instance. Consequently most of the searches are unknown item search.

An unknown item search is the one where users are not aware of the existence of any document that may solve their problems. In other words, users do not know whether or not any such item exists that can meet their information requirements.

Exact Match Search

In exact match search, the search engine will only match query terms exactly; it does not allow for truncation, wildcards, or stemming. Exact Match option is nowadays available in Internet-based databases to retrieve more relevant information. Phrase search can be characterized as exact match search, where a phrase is given at the search query that searches whole phrase.

Best Match Search

In best match search, the search engine will match query terms closely, if not exactly. It may allow for truncation, wildcards, or stemming. Best Match search is performed, when exact match could not fetch sufficient number of relevant information. Best match search constructs a tree-structured self-organizing map, where each level of the tree consists of a separate, progressively larger self-organising map. The search for the best match then proceeds level by level, at each time restricting the search to a subset of units that is governed by the location of the best match in the previous, smaller level. The map is taught one level at a time, starting from the smallest level. The best match search can be done even more quickly if the data set is relatively small: the location of the best match in the previous level can be tabulated for each input sample.

Partial Match Search

A partial match is one that matches one or more characters at the end of the text input, but did not match all of the regular expression, although, it may have done so had more input been available. Partial matches are typically used when either validating data input, checking each character as it is entered on the keyboard, or when searching texts that are either too long to load into memory or even into a memory mapped file, or are of indeterminate length, for example the source may be a socket. Some information retrieval systems perform partial match search.

Information Seeking and user Interfaces

The user interface forms an important component of an information retrieval system since it connects the users to the organised information resources. A user interface is the means by which information is transferred between the user and the computer and vice-versa. Well-designed user interfaces should allow the users to better find and fully use the information that the information system provides access

NOTES

to. In fact a good user interface greatly enhances the quality of interactions with information systems.

User interfaces basically perform two major functions: (*a*) they allow users to search or browse an information collection, and (*b*) they display the results of a search, and often allows users to perform further tasks, like sorting, saving and/or printing the search results, modifying the search query, and so on. The user interface therefore is the most important component of an information retrieval system that a user can see and interact with. The success of an information retrieval system depends significantly on the design and usefulness of the user interface. Hence, significant amount of research has taken place in the past few decades on the design, use and evaluation of user interfaces to various kinds of information retrieval systems.

Information Need and Information Seeking

The user is the focal point of all information retrieval systems because the sole objective of any information storage and retrieval system is to transfer information from the source (the database) to the user. Information need is often a vague concept. It is often a result of some unresolved problem(s). It may arise when an individual recognizes that his/her current state of knowledge is insufficient to cope with the task in hand, or to resolve conflicts in a subject area, or to fill a void in some area of knowledge. Information, needed by the user to accomplish a goal – to resolve a problem, to answer a specific question, or to meet a curiosity—may vary from quick and brief information to the most exhaustive and detailed information.

Although it appears to be a very simple model, in essence several complex processes take place throughout the process. Some of these processes are technological and are related to the information retrieval system, users interfaces, etc. Other processes relate to the nature and characteristics of the content as well as the concerned user. The process may take more or less time, and may become simple or complex depending on the nature of the users – their cognitive abilities, background, specific nature of the information need, and so on.

Features of Information Retrieval Systems

Based on accessibility of information, two broad categories of information retrieval systems can be identified: in-house and online. In-house information retrieval systems are set up by a particular library or information centre to serve mainly the users within the organization. One particular type of in-house database is the library catalogue. Online Public Access Catalogues (OPACs) provide facilities for library users to carry out online catalogue searches, and then to check the availability of the item required.

By online information retrieval systems we mean those that have been designed to provide access to remote database(s) to a variety of users. Such services are available mostly on a commercial basis, and there are a number of vendors that handle this sort of service. With the development of optical storage technology, another type of information retrieval system appeared on CD-ROM (compact-disc read-only memory). Information retrieval systems based on CD-ROM technology are available mostly on a commercial basis, though there have been some free and in-house developments too. Basic techniques for search and retrieval of information from the in-house or CD-ROM and online information retrieval systems are more or less the same, except that the online system is linked to users at a distance through the electronic communication network.

Recent developments in computer and communication technologies have widened the scope of online information retrieval systems. The Internet and World Wide Web have made information available for use by anyone virtually anywhere with access to the appropriate equipment. This has led to the concept of a digital global library system where information can be generated and made available in electronic form on the Web for use by any user from any corner of the world. This of course involves a number of technical and management issues that need to be resolved in order to make the global digital library concept a reality.

Features of Different Types of Information Retrieval Systems

In today's digital library world, a user can get access to different types of information sources in a variety of formats. For example, a digital library may contain simple catalogues of information resources, like OPACs (Online Public Access Catalogues), or may contain full texts of documents, images, audio and video materials. The information resources may be available in different formats, and they may have been produced by using different types of hardware and software. For example, the text may be in MS-Word, or PDF, or in HTML format; images may be available in GIF or JPEG file formats. These information resources may reside on a number of different servers – local as well as remote – and they may have been indexed differently. All these issues make the information retrieval process very complex.

The following list represents the common choices that a user may have today from a digital library:

- OPACs
- Electronic databases
  - Online search services
  - CD-ROM databases

NOTES

- e-Journals
- Digital libraries
  - Local digital libraries
  - Remote digital libraries
- Web resources

Characteristics of the information retrieval systems that work behind all the above information channels or systems are discussed briefly in the following sections.

Information Retrieval Features of Online Search Services

Traditional online information search systems that began about four decades ago were designed to provide access to remote databases, often through a database vendor or service provider. These systems were expensive to use. They were not quite suitable for searching directly by the end-users, and in most cases were used by information intermediaries on behalf of, or in cooperation with, the end-users. Online search services have been provided by database producers, but more commonly by service providers or vendors like Dialog, Ovid, etc. The major characteristics of this type of online information retrieval system are as follows:

- users get access to remote databases that are often many in number and large in size;
- many databases can be searched using a single search interface;
- database records mainly contain bibliographic details of records with abstracts, and sometimes with additional information, such as citations, etc.; only some databases contain full text information;
- service providers have their own search interface with good search and retrieval capabilities;
- users need to register with the service providers;
- users are charged for searching as well as for the content; and
- modern online service providers have web interfaces with good search features and hyperlinked records/information.

Although each online search service provider, such as Dialog, Ovid, STN, etc., has its own proprietary retrieval engine and user interface, the commonly available search and retrieval features are as follows [*Chowdhury and Chowdhury, 2001a*]:

- Users can select one or more databases to search;
- Novice and expert search modes are available;

A search can be conducted with one or more keywords or phrases;

- Common search facilities include: Boolean search, truncation (some systems also allow users to search for the variant forms

of a word), proximity search, and field search (number of fields that can be searched depends on the chosen database);

- Searches can be limited by applying certain restrictions, such as language, date, type of material, etc.;
- A search can be conducted for a range of period (date of publication, for example);
- Some systems show the frequency of occurrence of the search terms in the output;
- Dialog provides a unique facility of searching through a common index file that allows users to select databases appropriate for a search topic;
- Some systems provide access to thesauri through the search interfaces; and
- Search results can be sorted and sometimes ranked by selected criteria.

Information Retrieval Features of OPACs

Online Public Access Catalogues (OPACs), though are quite different in terms of content, structure etc., from online databases, also provide access to remote databases. OPACs form an important part of a library's services. Features of OPACs can be summarised as follows:

- OPACs allow users to search for the bibliographic records contained within a library's collection;
- Nowadays, some OPACs also provide access to the electronic resources and databases, in addition to the typical bibliographic records;
- Searches take place on the metadata of the records in the library's collection;
- Sometimes users can search more than one collection (within the same library or in different libraries);
- They have relatively simple search interface; and
- OPACs are nowadays available through the web.

Although each OPAC has a search interface and retrieval engine that is proprietary to the company providing the software for the purpose, the following information retrieval features are commonly available in OPACs:

- Browse and search facilities;
- Keyword and phrase search facilities;
- Indexers assign subject headings to the records by using a subject heading list like LCSH (Library of Congress Subject Heading List), and users can search by these assigned subject headings;

- Boolean search usually limited to the keyword search option; in other words, only keywords can be combined with Boolean operators;
- Proximity search also limited to the keyword search option;
- Search results are usually not ranked;
- Searching of records through selected keys – author, title, ISBN, call number, etc.; these are searched as phrases, and are usually automatically right truncated; and
- Some searches can be limited by date, collection, language, etc.

3.3 FILE ORGANISATION IN ISAR SYSTEMS

Information Storage and Retrieval (ISAR) system deals with three basic aspects:

- Information representation
- Information storage and organisation
- Information access.

One of the best examples of ISAR system is library system, where information is stored, processed, organised and retrieved on demand. Information could be stored in a book, audio-video, images and so on. The library and information centres endeavours to organize knowledge available in documentary form. The multi-faceted universe of knowledge is represented in libraries in linear form using some classification scheme. At the time of retrieval, specific aids like cataloguing and indexing are used and meaningful information is retrieved.

A lot of emphasis is given on improving the performance of the system. For that librarians have developed classification schemes for helpful arrangement of documents on the shelf so that retrieval can be facilitated. The tools like library catalogue or indexes are developed and further modified to satisfy the different approaches of users toward information. Automated systems reduce the effective time of users in searching for information which, in effect, further improves the performance of the system. Therefore, tools like Database Management System (DBMS) are used for keeping records of a holding of library, which is known as Online Public Access Catalogue (OPAC).

Objectives of ISAR Systems

The principle objective of ISAR system is to provide correct information to the user in least time with least efforts. Thus, while designing any ISAR a system designer should keep following objectives in view:

NOTES

Information Facilitator

The ISAR system should act as facilitator between the information (contained in document) and the users. If a user approaches with the subject term, name of contributors or title of the document and so on, the system should be helpful to give him the desired information. The information could be exact information or the reference of a document which contains information.

Non-Ambiguous

The system should be so organized that ambiguity of information is avoided so that search result is free from any kind of ambiguity. This requires identification of terms, setting their context and their proper indexing. For example, search for a term 'screw driver' should not bring results like 'truck driver', 'hardware driver' and so on.

Minimum Time

The system should be so designed that minimum effort and time are spent to interrogate the system.

Searching through the system should take minimum time, meaning thereby that the ISAR should be capable of performing fast search. Not only that, it is best to have an online ISAR so that users do not need to walk to library. They should get whatever they want at their work place.

User Friendliness

Ease of use is an important consideration for any ISAR system.

Any ISAR should have user friendly interface. The important aspects of ISAR should be highlighted. Before a user uses the system he/she should be properly introduced to the system with all its features, *i.e.,* informing users about the scope of system, available search options, and most importantly how to perform search with the system. It is only this interface through which a user operates an ISAR system. Take an example of a Library OPAC. It should have following features:

- Introduction to library
- Scope of collection
- Instructions for performing search

The search interface should facilitate framing the search like,

- Keyword search
- Author and title search
- Combination search (using Boolean operators)
- Proximity search, etc.

While designing an ISAR system the following aspects should be recognised to achieve the objectives of the system:

NOTES

- The desirability of making systems as readily usable as possible for their clienteles;
- The need to recognise basic features of retrieval system; and To incorporate coordinating features such as vocabulary control, search strategies, user-interface, information modelling aspects in general, etc.

Features of ISAR System

Keeping in view the objectives mentioned above and recognizing the aspects to be considered in designing a system, an ideal ISAR system should incorporate one or more of the following features:

- The competence and compatibility for consolidated searching and retrieval of information from any client terminal from any database within the system.
- It should be able to narrowcast or broadcast or relate the information need in a variety of associations to get optimum retrieval performance.
- It should have access facilities at multi-points.
- It should have common command language facility to retrieve information from several databases of the system.
- It should be able to handle information access from entity-related or object-oriented approaches. It may also provide all other associations for accessing information.
- In a bibliographic or full-text database, the surrogates chosen should have indicative as well as informative features that are sufficient enough to select or reject the retrieving information based on end-users' needs.
- It should have the ability to select, classify, process and consolidate the analysed information into a cohesive text ready for assimilation by the end-users.
- It should have ability to orient the information to specialist needs of the users from time to time. This calls for understanding the processing of user profiles.
- It should be able to retrieve maximum information with minimum number of clues.
- The fuzzy approaches of end-users must be able to get clarified and ultimate result should provide satisfaction to the searcher.
- It should have capacity to interchange the information available in one database or another for purposes of retrieval relevance end usage.
- It should have bibliographic data interchange capacity (using Z39.50 or similar standard) to meet consolidation to a chosen format for networking and other purposes. Compatibility with standards at all levels must be the goal.

- It should have ability to search simple information quickly in an easy manner and also have the ability to multi-track the complex questions and present them in a simple easy manner. User-friendly presentations are very important.

Types of ISAR Systems

ISAR systems are used by a wide range of users. According to different kinds of needs and purpose of use, different types of automated systems may be designed. Such types may be:

- Database Management System (DBMS)
- Text Retrieval System
- Management Information System (MIS)
- Decision Support System (DSS)
- Knowledge Based System (KBS)

Database Management System (DBMS)

Any automated system is based on a collection of stored information or documents in a database which is amenable for access. A DBMS is primarily concerned with data storage, maintenance and retrieval and is used to keep control and manipulate data within the database.

The distinguishing characteristic in DBMS is the definite structure of the stored information, instead of dealing with natural language text. In DBMS, normally files of data are described by a small set of pre-specified attributes. For example, in a Salary database of an organization, name of person, designation, salary, etc.. are attributes. Similarly, in the context of records of books author, title, publisher, year etc., may be the attributes. Each attribute carries some kind of value in it. Therefore, a DBMS can be defined as set of records and each record contains fields (attributes) which in turn contain data (value). A database may contain textual, numeric, statistical and graphical information.

OPAC is a kind of DBMS often built of some kind of Bibliographic Database Management Systems (BDBMS). The typical example of BDBMS is the one built by CDS-ISIS/WINISIS developed by UNESCO. The data fields may contain author, title, place, publisher, year of publication and so on.

Text Retrieval System

In contrast to DBMS, text retrieval systems are designed for unstructured data such as full text documents. Queries are usually language based here such as, keywords and a number of advanced search techniques (such as proximity search) can be used. However, systems may also handle discrete structured data.

NOTES

Management Information System (MIS)

Management Information System is a kind of database management system designed to cope up with the needs of managements who need to have information about different alternatives related to his/her interest to facilitate his work. Though built on DBMS platform, information are subjected to special processing. In such a system information is available with different alternatives. In the business environment managers needs to take complex and rapid decisions. Under such circumstance MIS is a kind of sophisticated tool which provides them timely information to take decisions.

Collection of data is critical in MIS because the information comes from different sources, *i.e.*, within the organization or from outside organization. Collection of data not only needs defining that how the data would be captured but also the estimation of cost involved in data collection. Once the data are collected and organized such system generates reports for usage. The reports could be generated in printed or electronic form depending on desired format. Such systems also generate reports upon different intervals if it is desired.

Decision Support System (DSS)

Decision Support systems help top-level management in arriving at decisions. There is very little difference between MIS and DSS. The former generates reports in anticipation or on demand and collect facts, whereas the later provides possible alternative solutions. These are interactive computer-based systems that provide the user easy access to decision models and data in order to support semi-structured and unstructured decisions. In management parlance a structured decision means use of rules and norms for making decision. Such systems help managers in identifying the problem, analyzing alternatives and choosing possible solution. However, the typical decision-making takes place with shared effort of human and machine. In true sense, a DSS cannot take decision, rather it amplifies decision maker's capability by providing resources and facilities. In other words it provides intellectual support. These systems are integrated with powerful tools like generating charts, preparing tables and presentation tools.

Knowledge Based System (KBS)

Specialised computer programs, modeled in the same way as human experts tackle problems and arrive at solutions, are called 'Expert Systems'. Such systems rely upon a store of specialised knowledge for solving problems and hence referred to as Knowledge Based Computer Systems (KBCS) or Knowledge Based Systems (KBS).

Expert systems are sophisticated computer programs that manipulate knowledge to solve problems efficiently in a narrow problem area. Knowledge based systems enhance the value of expert knowledge by making it readily and widely accessible. Like human experts, these

systems use symbolic logic and heuristics to find solutions. They are also capable of learning from experience through inferencing mechanism.

KBS systems are domain specific and are backed up by a strong knowledge base. In these systems each bit of information is not only stored but they are also linked. This linking is used to preserve the context. The context is used to draw the inference from a query. They are capable of providing solutions and replace the human intervention. Expert systems are thus data driven. But it is also important that how the given information is utilized to achieve the goal. Human experts possess procedural knowledge (How To), which helps them to flowchart the courses of action to be taken in solving problems. Accordingly, rules for manipulation of the knowledge have to be incorporated in the expert systems to get the desirable results. However, this procedure involves other related tasks such as intuitive inferencing power, learning and updation of knowledge. Solutions to problems can be achieved if only all tasks are executed in a coordinated way.

Design of ISAR Systems

From the foregoing discussion, it should be clear that a number of features should be ingrained in an ISAR system. Thus, an ISAR system is an integrated one combining various aspects.

Components of an ISAR System

An ISAR system comprises of following components:

- User Interface
- Knowledge Base
- Search Agent

The features to be incorporated in each components are discussed below:

User Interface

User interface is the part which puts users across the ISAR system. It is the front-end which enables user to put a query and displays results. Basically, user interface is of two types:

- Query Interface
- Result Interface

Query Interface

This is the end from where users enter his/her search terms. It is one of the major components which initiate communication between users and the system.

The Query Interface should have following features:

(*a*) **Understanding the user Input Statement.** It is also known as front-end. The interface should be able to capture keywords given by users which should be passed on to the search program.

NOTES

The front-end should have understandable look and feel. At the time of designing the system, one should also seriously consider use of different colours, and instructions for performing search and limitation of search.

(*b*) **Refining the Problem Statement.** Sometimes users start with broader domain and further refine the search. The interface should have flexibility for further refining any query within the displayed search results. In other words, system should provide facilities for further modification of search statement. It should also display some kind of arrangement among topical terms which further facilitate browsing through the system.

(*c*) **Search Statement to Search Strategy Translation.** Any knowledge base accepts a query in a particular format. For example, a Relational Database Management System (RDBMS) accepts search statement in Structured Query Language (SQL). It is the system front-end which translates a search statement and formulates a search strategy in the language which is understood by Search Agent.

(*d*) **Modification of Search Strategy.** If one does not get desired output from the database, ISAR system should have procedure for further modification of search strategy. The modification should be interactive. Vocabulary control devices can also be added as an aid for users to locate the term of his/her interest.

Result Interface

Display of search results is another important aspect of searching. It should be in user friendly manner. Not only that the result should cater the needs of individual users but the display should also be customized. Search results should also display the ratings in the light of search terms. For this purpose statistical techniques can be used.

Knowledge Base

The store house of any ISAR system is its Knowledge Base. It contains list of facts or related facts (information). Any kind of query is answered based on the facts stored in the Knowledge Base. A Knowledge Base could be a Database Management System (DBMS). Retrieval of information from storage depends on two important aspects of Knowledge Base:

- Knowledge Representation
- Indexing and Clustering

Knowledge Representation

The first and foremost objective in constructing an ISAR system is representation of facts within the Knowledge Base. There are different ways of representation of knowledge:

NOTES

Semantic Network Knowledge Representation

Semantic network is a method of knowledge representation based on a network structure. A semantic network contains points called nodes connected by links called as arcs. The nodes represent objects, concepts or events — in other words documents or information. The arcs are used to represent the relations between the nodes. Arcs build a kind of hierarchies in the Knowledge Base. Arcs usually represent relations like is_a or has_part. For example, Universe of Knowledge à Library Science à Cataloguing

Semantic networks are a useful way to represent knowledge in domains that use well established taxonomies to simplify problem solving. Semantic networks are useful in representation of sentences of natural language.

Frame Based Knowledge Representation

It is an object-oriented approach. A frame represents an object (document or information) or class of objects (collection of documents or information) or several facts. When they represent a class of objects, they generalise certain groups identifying overall properties of those groups, it shares. The pointers where properties are stored are known as slots. Similarly, if frame represents an object, slots represent the properties or attributes of the object. Slots contain value for that particular attribute. For example, a book in a library is an object, therefore it can be represented as frame. The properties of book, *i.e.*, Title, Author, Place, Publisher and so on are stored as slots and each slot would have corresponding value.

Frames are also very helpful in representing hierarchies. In a frame base, slots can store procedures or relations. Relations are used for storing taxonomy or genealogical data or parent child information. For example, if a person is represented by a frame, different slots can store value who is the father or who is the son of the person. Similarly, slots can also store procedures that means how a frame or object should react or the way it should be used or it can operate with other objects.

Rule-Based Knowledge Representation

Rule based representation is a popular approach. Rules are employed to state the way in which the inferencing has to be done. Rules provide a formal way of representing recommendations, directives, or strategies. Rules are appropriate when the domain knowledge results from empirical associations developed through years of experience in solving problems in a given area. Rules are expressed in the form of IF-THEN statements.

In a rule based expert system, the domain knowledge is represented as a set of rules that are checked against a collection of facts or knowledge about the current situation. When the IF portion of the rule is satisfied by the facts, the action specified by the THEN portion is performed.

NOTES

When the condition is satisfied the rule is said to 'fire' or 'execute'. A rule interpreter is used to compare the IF portions of rules with the facts and execute the rule whose IF portion matches the facts.

Indexing

Many of the systems follow a kind of keyword indexing. Unfortunately the keyword indexes are sometimes good in recall not in precision and in some cases vice-versa. Such indexes fail in preserving the context of search term. WWW is a good example of use of keyword indexing. There has been different efforts made in this direction such as using clustering techniques like keyword clustering to attach semantics with a keyword. In such a technique, relation among the terms are used like Broader term, Narrower term and Related term. This technique heavily uses Thesaurus. Such relations can be stored in the form of sequential inverted files or using B-tree structures.

Search Agent

Any ISAR system should be backed up by a Search Agent. It is a program which takes input from Search Interface and searches in the Knowledge Base using existing index. A good ISAR system means efficient retrieval. Thus, a good search agent must be equipped with following features:

- facility of using Boolean operators
- context setting to search terms
- use of clustering algorithms
- use of soundex and metaphone algorithms

Boolean Operators. Three Boolean operators are AND, OR and NOT. These operators are used to generate combinational search. AND and NOT operators increase precision whereas OR increases recall of search results.

Context Setting. Context Setting requires content analysis of document. Here one analyses document manually or automatically in order to preserve the context of each term in the index. It can be done in two ways, *i.e.*, Conceptual Analysis and Relational Analysis. Conceptual analysis can be thought of as frequency of concepts. Concept can be represented by texts as well as pictures where text is very common. To analyze the concept one looks for the appearance of words in the text. It is not necessary that same word appears always, there may be synonymous terms present. For example, if one is analyzing with a hypothesis that a certain document is about freedom then one should look for the related words like liberation, independence, etc. In contrast, relational analysis goes one step further by examining the relationships among concepts in a text. In relational analysis we look for what are the related words appearing next to the word in question. For example, to see what are the words that appear next to freedom and then determine the related concepts.

Clustering Algorithms. Clustering is a method by which large sets of data is grouped into groups or clusters of smaller sets of similar databased on some characteristics. For example, in a group of players one can cluster players according to their specialisation of game, like those who play cricket, those who play hockey and so on. A clustering algorithm attempts to identify natural groups of components or data based on some similarity in a given population. In other words, it is a method to create subclass in a given class. The first thing in such algorithms are identification of core entity which is also known as centroid. Around centroid similar kind of entities or data are collected which are called as members of cluster. To determine cluster membership and size, most algorithms evaluate the distance between each entity or data and the cluster centroids. Statistical techniques are used for generating clusters as output.

Soundex and Metaphone Algorithms. Soundex and metaphone algorithms are almost the same kind of algorithm. Both these algorithms are based in the way pronunciation of a word is made. In soundex algorithm, a numeric code is assigned to each character used in a word and when search is performed, words with similar codes are also brought out in search result. Metaphone is also same kind of algorithm but unlike soundex which encodes on letter-by-letter basis, it encodes groups of letters *i.e.*, a word. Metaphone embodies more accurately the rules of pronunciation in language.

Such algorithms are well established for English as a language. Both algorithms return all the words that exactly match the desired word as well as all similar sounding names.

Functionality of an ISAR System

The objective of an ISAR system is storage, processing and retrieval of information. In order to perform the above mentioned tasks, such systems are equipped with user friendly interface, powerful database management system and search agents. Information is represented inside the system in a machine readable format using frames or semantic networks or as rule base. Often such systems are combination of above three or two methods. The overall objective of the system depends upon the need of users. Once needs of users are defined, it is easy to frame a system for the purpose.

System stores information in a form of structured record which in turn is indexed using automated or manual technique. Often the search engines over Internet use Keyword indexes using automated programs called Robots or Spiders. It goes through various web pages on Internet and indexes them word-by-word. This index is stored inside databases similar to library catalogues. Any kind of search query is searched through the index created by Robots.

From the perspective of users, he/she approaches to ISAR system with some query in his/her mind. Then he/she transfers the query

NOTES

into a search strategy with fixed vocabulary or key term or in a fixed format (for example, SQL query). Once query is made, he/she is given a list of search results out of which one chooses the best ones of his/her interest.

At this level a user goes through the search results and evaluates the displayed result in the light of his/her query. If the result displayed does not suit to his/her needs then he/she can go for further refinement of search strategy.

Steps in the Development of ISAR Systems

The development of an ISAR system starts with the recognition of the need for developing a system in relation to the needs of the users, as all the subsequent activities are dependent upon these. When designing, ISAR systems should follow System Development Life Cycle (SDLC) for greater efficiency and effectiveness of the systems. In the design of the systems, following steps need to be followed :

(*a*) Recognising the need for development of ISAR system.

(*b*) Recognising the information needs of the users.

(*c*) Identification of users need.

(*d*) Deciding the type(s) of databases to be incorporated into the system.

(*e*) Deciding about the features for incorporation in the databases.

(*f*) Preparation of structured queries.

(*g*) Design and development of various components of the system such as user interface, search agent, etc.

(*h*) Evaluation of the system.

(*i*) Re-designing/Modification of ISAR system, if needed.

3.4 EVALUATION OF ISAR SYSTEMS

In an ISAR system search can be performed in manual or automated environment. An example of manual system is library shelf-list or card catalogue whereas OPAC or Digital Libraries are examples of automated systems. According to Claverdon and others any ISAR system can be evaluated based on the following points:

- The coverage of the collection, that is, the extent to which the system includes relevant matter;
- The time lag, that is, the average interval between the time the search request is made and the time an answer is given;
- The form of presentation of the output;
- The effort involved on the part of the user in obtaining answers to his search requests;

- The recall of the system, that is, the proportion of relevant material actually retrieved in answer to a search request; and
- The precision of the system, that is, the proportion of retrieved material that is actually relevant.

If the view of Claverdon is taken and further analyzed one gets following check points, on which any ISAR can be evaluated:

- Coverage
- Cost Benefit Analysis
- Time

The important issue which should be taken into consideration while evaluating an ISAR, is the Time factor. The first and foremost thing which should be observed, *How long a user takes to get the satisfactory* answer to his question? If a user spends more time over system searching for his/her answer, it means that the system has to be modified in such a way that one gets his/her answer immediately. Searching over an ISAR system depends a lot and varies from on person to person. Therefore such kind of measurement should be supported by a long observation.

Other evaluation criteria are:

- Completeness and Relevance
- Novelty Ratio
- Noise

The evaluation parameters discussed therein in the context of indexing systems are equally applicable for ISAR System as a whole or a sub-system of it.

Critical Aspects of Recall and Precision

There is a problem of determination of the total number of relevant documents in collections unless one scans the whole collections completely, which is almost unfeasible. Another thing is that relevancy cannot be measured only by yes/no terms, but also by the degree of relevancy. Recall and precision have inverse-relationship, where both cannot achieve the highest degree at a point of time. The recall and precision are influenced by the following factors:

- queries that imperfectly represent information needs;
- indexing factors;
- search strategy factors; and
- vocabulary factors.

Lack of specificity, lack of exhaustivity, lack of specific terms, inadequate hierarchical cross-reference structure, defects in hierarchy, failure to cover all reasonable approaches to retrieval, etc., are the major hindrance to achieve higher degree of relevant retrieval from an ISAR system which also affect recall and precision ratios.

NOTES

Systems Criteria for Evaluation

The basis for evaluation of an ISAR system could be done at every stage with every component of it. Once again it centres around the users, information sources, intermediaries, the tools, techniques, methodologies and overall environment. B.C. Vickery and Alina Vickery have provided a detailed structure for it.

After setting a framework for evaluation of ISAR system, they present a panorama and yet analyse view of ISAR system which is on the basis of quality and value as beneficial aspects of the evaluation system. User demands and their satisfaction through provision of right information will be ultimate aim of any evaluation.

Nowadays advanced technologies are used in modern ISAR systems. The full-text and multi-media contents are now common in modern ISAR systems. Also hybrid systems co-exist that provide access to physical collections as well as electronic collections. The criteria for evaluation, as suggested by B.C. Vickery and Alina Vickery, are still relevant in modern ISAR systems, but require some modifications to cope up with new techniques, new technologies and new types of documents.

3.5 EXAMPLES OF ISAR SYSTEMS

Online Public Access Catalogue (OPAC)

Library catalogue has changed its form to the electronic form with the advent of computerisation and today it is called the OPAC (Online Public Access Catalogue). OPAC provides access to the documents by different approaches of users, such as Author, Collaborator, Title, Subject, Keyword, etc. An OPAC is built on Data Base Management System.

An OPAC should provide facilities for searches like Boolean, truncation and proximity searches so that users can locate document without much effort. OPAC uses bibliographic standards like MARC21, CCF (Common Communication Format), UNIMARC etc.

Digital Library

Online Public Access Catalogues (OPAC) provides only bibliographic details not full-text information. This limitation led to development of full-text databases in digital form. This kind of database is known as Digital Library. Digital libraries are available over network often using WWW.

The costs of creating, storing, and transmitting digital information have been decreased, and the technology to support distribution and access is widespread. Rising acquisition and subscription fees (not to mention shelving and processing costs) have forced libraries to seek other ways to make information available.

Perhaps more importantly, digital libraries support service improvement. Information search and navigation across electronic information resources are faster, with enriched points of access and alternative methods for browsing and exploration. The resources themselves can be segmented, rearranged, annotated, and enhanced in ways not possible before. Digital libraries provide remote access to information resources to the users on their desktops. Also, digitization of documents presents opportunities for long-term preservation of bodies of knowledge, if not of the original carriers of that knowledge.

Search Engines

WWW is itself is a big source of information. Almost everything and anything can be found over Internet. Search Engines provide a kind of interface for users to search the web. A Search Engine basically has three components:

- a Robot or Web Crawler
- a Database
- an Agent

Web Crawler goes to each and every site over Internet and indexes each word present in the page or sometimes few lines from the page. This index is stored in search engines database with corresponding URL (Uniform Resource Locator). When a search query is given it searches in databases of search engine and result is generated.

Search Engines can be categorised into three main types:

- Individual Search Engines – text or image based search engine, for example, Google (www.google.com)
- Subject Directories – subject-based search engines, for example, Yahoo! (www.yahoo.com)
- Meta-search Engines – search engine of search engines, for example, Askjeeves (www.askjeeves.com)

3.6 SUMMARY

- Libraries are store house of information. In a big collection, it is very difficult to locate a document or any piece of information in the library. For this purpose libraries had evolved tools like catalogue and index. With the advent of Information Technology and development of automation of libraries the tools like catalogue and index are being transformed into Automated Information Storage and Retrieval (ISAR) Systems. A typical ISAR system has following components:
  - User interface
  - Knowledge base
  - Search agent

NOTES

- In course of time and with the expansion of IT, several applications of ISAR has been developed. The most important product of automated ISAR system is Digital Libraries.

3.7 REVIEW QUESTIONS

1. Mention some synonyms of information retrieval.
2. What is a database? Give examples of three bibliographic databases.
3. What is an inverted file? What role does it play in an information retrieval process?
4. What is the role of a vocabulary control tool in an information retrieval process?
5. What is the difference between a subject heading list and a thesaurus from the perspectives of information retrieval?
6. What is a user interface?
7. What are the two major functions of the user interface in an information retrieval system?
8. What are the objectives which should be kept in mind while designing an ISAR?

3.8 FURTHER READINGS

1. Aitchison, J. and Gilchrist, A. (2000). Thesaurus construction and use: a practical manual. 4th ed. London: Aslib.
2. Belkin, N.J. (1980). Anomalous states of knowledge as a basis for information retrieval. Canadian Journal of Information Science, 5, 133–43.
3. Brooks, H.M. (1987). Expert systems and intelligent information retrieval. Information Processing and Management, 23 (4), 367–82.
4. BS 5723: 1987. Guidelines for the establishment and development of monolingual thesauri. London: British Standards Institution.
5. BS 6723: 1985. Guidelines for the establishment and development of multilingual thesauri. London: British Standards Institution.
6. Chowdhury, G.G. (2004). Introduction to modern information retrieval. 2nd ed. London: Facet Publishing.

UNIT IV INFORMATION ACCESS AND RETRIEVAL

NOTES

★ STRUCTURE ★

LEARNING OBJECTIVES

After going through this unit, you will be able to:

- describe information retrieval
- know about database and information base
- explain structured query language

4.1 INTRODUCTION

Pattern of information retrieval indicates a knowledge seeking behaviour of individuals and groups of individuals. In this context, a searcher seeks some information from the vast store of a knowledge. An analysis and diagnosis of this state of mind provides guidelines for organisation of knowledge in libraries, information retrieval systems, databases, knowledge bases and similar environments. Such guidelines are aimed at providing conducive compatibility between searchers' approach and knowledge organisation in a database. The current human environment consisting of learning, problem-solving, and decision-making situation calls for flexibility in knowledge structures at each instance. The development in computer and communication technologies have made it possible to store vast amount of information in compact form. The variety of software developed have also given scope for quick retrieval of information from this store. While the speed of retrieval is valuable,

NOTES

it could be enriched further if the retrieved information is readily assimilable by the information seeker. It is in this context that modeling of retrieved information into user-friendly approaches calls for cognitive modelling of information retrieval.

Such development has given rise to a field **'Cognitive Science'** which is an inter-disciplinary field drawing inputs from the fields of Psychology, Behavioural Studies, Computer Science, Artificial Intelligence and Information Science.

4.2 INFORMATION RETRIEVAL

Information Retrieval is a process of selecting information from a store. It connotes that search for information may be from documents, metadata which describes documents or searching interim databases. The databases may be standalone databases or hypertext networked databases like Intranet and Internet. It primarily helps a person who needs to get some information in his activities, be it research, problem solving, decision-making, production, service, etc. Broadly speaking, four kinds of information retrieval exist. Document retrieval, in which simple structured files are normally processed, using a small number of well-defined attributes to characterise each record, and a restricted set of pre-specified query types to access the database; reference retrieval, in which the records represent books, documents, and other library materials, and the number of different attributes available for the identification of the information items is effectively unlimited. In that case, the queries often refer to the information content of individual documents. In the most general case, a retrieval system might be designed to handle any kind of query and the system might furnish direct replies to such queries; in fact retrieval, a wide variety of different types of information identifiers may be needed, and the answers may have to be based not only on a deep analysis of each individual information item, but also on general world knowledge and other extraneous factors. In text retrieval, instead of retrieving reference or data or surrogates of documents, text on a particular topic are retrieved. Irrespective of any retrieval environment, the following four main system components must be taken into account in formulation of the retrieval problem.

(*a*) The objects, documents, or records themselves (which in the aggregate constitute the information files to be processed);

(*b*) The information identifiers, terms, index terms, keywords, attributes, etc. (which characterise the records or documents and represent the information content in each case);

(*c*) The information requests (which enter into the system and are to be compared with the stored records for retrieval); and

(*d*) The relevance information (often supplied by the users of the system connecting the information requests to the stored information items).

NOTES

An Information Retrieval System is a system that is capable of storage, retrieval, and maintenance of information. Information in this context can be composed of text (including numeric and date data), images, audio, video and other multimedia objects. An Information Retrieval System thus consists of a software program that facilitates a user in finding the information of his needs. The system may use standard computer hardware or specialized hardware to support the search sub-function and to convert non-textual sources to a searchable media (*e.g.*, transcription of audio to text). The success of an information system is gauged by how well it can minimize the overhead for a user to find the needed information. Overhead can be expressed as the time a user spends in all of the steps leading to reading an item containing the needed information (*e.g.*, query generation, query execution, scanning results of query to select items to read, reading non-relevant items). The success of an information system is very subjective, based upon what information is needed and the willingness of a user to accept overhead. Under some circumstances, needed information can be defined as all information that is in the system that relates to a user's need. Thus, search composition, search execution, and reading non-relevant items are all aspects of information retrieval overhead.

The first information retrieval system originated with the need to organize information in libraries. Catalogues were created to facilitate the identification and retrieval of items. The term 'item' is used to represent the smallest complete unit that is processed and manipulated by the system. The definition of item varies by how a specific source treats information. A complete document, such as a book, newspaper or magazine could be an item. Each chapter, or article may also be defined as an item. As sources vary and systems include more complex processing, an item may address even lower levels of abstraction such as a contiguous passage of text or a paragraph.

With the advent of inexpensive powerful personal computer processing systems and high speed, large capacity secondary storage products, it has become commercially feasible to provide large textual information databases for the average user.

In information retrieval the term 'relevant' item is used to represent an item containing the needed information. From a user's perspective 'relevant' and 'needed' are synonymous. From a system perspective, information could be relevant to a search statement (*i.e.*, matching the criteria of the search statement) even though it is not needed, relevant to user (*e.g.*, the user already knew the information).

In addition to the complexities in generating a query, quite often the user is not an expert in the area that is being searched and lacks

domain specific vocabulary that is unique to that particular subject area. The user starts the search process with a general concept of the information required, but does not have a focused definition of exactly what is needed. A limited knowledge of the vocabulary associated with a particular area along with lack of focus on exactly what information is needed leads to use of inaccurate and in some cases misleading search terms. Even when the user is an expert in the area being searched, the ability to select the proper search terms is constrained by lack of knowledge of the author's vocabulary.

There are natural obstacles to specification of the information, a user needs that come from ambiguities inherent in languages, limits to the user's ability to express what information is needed and differences between the user's vocabulary corpus and that of the authors of the items in the database. Languages suffer from word ambiguities such as homographs and use of acronyms that allow the same word to have multiple meanings. Many users have trouble in generating a good search statement. The typical user does not have significant experience with, nor even the aptitude for Boolean logic statements. The use of Boolean logic is a legacy from the evolution of database management systems. Multimedia also adds an additional level of complexity in search specification.

Thus, an information retrieval system must provide tools to help to overcome the search specification problems. In particular, the search tools must assist the user automatically and through system interaction in developing a search specification that represents the need of the user and the writing style of diverse authors.

There are three basic components in a generalized modern computer-based information retrieval system. They are, the database; the information seeker; and the retrieval techniques, tools, models, and processes which attempt to bridge the gap between searcher and the database.

4.3 DATABASE

It is designed to take-in information and information sources acquired for the specific purpose of serving user groups. The selection of information and other sources is based on this objective of serving them. A database has a logical and physical organisation. It is an arrangement based on the users' approach. The main purpose of a bibliographic database is to collect and collate standard bibliographic information, assign index terms and abstracts for technical papers and monographs, etc. In case of full text databases, the complete text of document is made available along with bibliographic details.

Information Retrieval Systems are increasingly using unformatted, free form means of storing information. Record-keeping files in a traditional paper based system might have, for example, a separate line for a person's name, a separate line for an address, a place for a seven-digit phone number, and so on. In computer systems, it is now more common to have information stored in a less specifically located form, making more difficult to locate specific features such as a name or an address. The increased use of word processing, the proliferation of desktop personal computers, and the introduction of optical scanners that can read and convert material into organizational databases, which store information in a way that can make it more difficult to recreate structurally, are making information storage systems more popular.

On the other hand, precisely because information can be stored in a free form, unformatted way, electronic means of storage also make it possible to retrieve and manipulate information in ways that are not possible in a paper-based system, as our increasing reliance on electronic means of storage and retrieval system makes it clear. The difference in power and flexibility in retrieving information can be illustrated by comparing searches done in print indexes and their electronic counterparts. A search in a print index restricts the user to chosen keywords or index terms that can be combined in relatively limited ways. A search in an electronic database generally allows freedom to the searcher and allows him to define more precisely the hierarchical and proximal relationships the terms should have.

4.4 INFORMATION BASE

An Information Retrieval System locates and presents information to the user, based on a query presented to the system. A query may indicate precisely the characteristics of information to be retrieved, or it may express as approximation of information need, indicating merely an initial guess as to the characteristics of the information to be retrieved. A precise-query is most commonly found where an answer is needed to a single factual question, such as "which is the capital of Gujarat State". Question requiring more complex or more ambiguous answers may be expressed by queries that are less precise. Take for instance, an organization that desires to hire internally to establish an information system that will store and retrieve documents in multiple languages. A secretary in a personnel department would pose a question, such as, "who has the experience with multilingual systems". This question is looking for a series of potential answers. Information Retrieval System must be able to receive different kinds of queries and answers in different ways, retrieving individual facts or groups of potentially relevant items.

Such flexibility is necessary because of the widely varying sources and reasons for needing information, which are rooted in what Belkin, Brooke and Oddy [1972] refer to as Anomalous State of Knowledge (ASK). These ASKs may represent the lack of a particular fact in the information seeker's knowledge base, they may represent a much larger area of missing information, or they may represent a lack of knowledge structure. A small need can often be answered with a single fact, while a much larger or more unstructured area within one's knowledge base may potentially be answered in a variety of ways be a number of facts or documents. Information base, whether databases or collections, must be structured and organised to meet the potential needs of the people who will use them, so that the means of information retrieval must also be selected or created with an understanding of what information needs will have to be met, and how people are likely to understand and use the system.

4.5 STRUCTURED QUERY LANGUAGE (SQL)

Structured Query Language (SQL) is a query language used for accessing and modifying information in a database. Some common SQL commands include 'insert,' 'update,' and 'delete'. Queries take the form of a command language that lets a user select, insert, update, find out the location of data, and so forth. There is also a programming interface. The language was first created by IBM in 1975 and was called SEQUEL for "Structured English Query Language". Since then, it has undergone a number of changes, with a lot of influence from Oracle Corporation. Today, SQL is commonly used for Web database development and management. Though SQL is now considered to be a standard language, there are still a number of variations of it, such as mSQL and mySQL. Many database products such as MS-Access, SQL Server and Oracle support SQL with proprietary extensions to the standard language. Some Information Retrieval Systems are limited to finding those facts or documents containing characteristics specified by the query. Such systems are often referred to as database systems or database as SQL or forms variant of it recognized by the system under consideration. SQL allows precise specification of the value for attributes of terms to be retrieved.

An example of an SQL query might be: Select from courses where Student's Name = "Anjali Kapoor" and Department = "Management Science".

4.6 SUMMARY

- The development in computer and communication technologies have made it possible to store vast amount of information in compact form.
- Context that modeling of retrieved information into user-friendly approaches calls for cognitive modelling of information retrieval.
- Information Retrieval is a process of selecting information from a store.
- An Information Retrieval System is a system that is capable of storage, retrieval, and maintenance of information.
- The success of an information system is gauged by how well it can minimize the overhead for a user to find the needed information.
- A complete document, such as a book, newspaper or magazine could be an item. Each chapter, or article may also be defined as an item.
- In information retrieval the term 'relevant' item is used to represent an item containing the needed information.
- There are natural obstacles to specification of the information, a user needs that come from ambiguities inherent in languages, limits to the user's ability to express what information is needed.
- An information retrieval system must provide tools to help to overcome the search specification problems.
- There are three basic components in a generalized modern computer-based information retrieval system.
- The main purpose of a bibliographic database is to collect and collate standard bibliographic information, assign index terms and abstracts for technical papers and monographs, etc.
- The increased use of word processing, the proliferation of desktop personal computers, and the introduction of optical scanners that can read and convert material into organizational databases.
- The difference in power and flexibility in retrieving information can be illustrated by comparing searches done in print indexes and their electronic counterparts.
- Information base, whether databases or collections, must be structured and organised to meet the potential needs of the people who will use them, so that the means of information retrieval must also be selected or created with an understanding of what information needs will have to be met.
- Structured Query Language (SQL) is a query language used for accessing and modifying information in a database.
- SQL is commonly used for Web database development and management.

NOTES

- Information Retrieval Systems are limited to finding those facts or documents containing characteristics specified by the query.

4.7 REVIEW QUESTIONS

1. Given example of Cognitive Science.
2. Differentiate between Intranet and Internet.
3. What do you mean by Information Retrieval?
4. How does success information is gauged?
5. Give the components of generalized modern computer based information retrieval system.
6. What do you mean by storing information?
7. For which purpose SQL is used?
8. Differentiate between mSQL and mySQL.

4.8 FURTHER READINGS

1. Stacey, Alison, Stacey, Adrian. (2004). Effective Information Retrieval from the Internet. Oxford: Chandoes Publishing.
2. Chowdhury, G.G. (2004). Introduction to Modern Information Retrieval. 2nd ed. London: Facet.
3. Fenichel, C.H. (1980). The Process of Searching Online Bibliographic Databases: A Review of Research. Library Research. 107–127.
4. Belkin, N.J. (1980). Anomalous states of knowledge as a basis for information retrieval. Canadian Journal of Information Science, 5, 133–43.
5. Brooks, H.M. (1987). Expert systems and intelligent information retrieval. Information Processing and Management, 23 (4), 267-82.

UNIT V DATABASE SEARCHING

NOTES

★ STRUCTURE ★

LEARNING OBJECTIVES

After going through this unit, you will be able to:

- know about database search service and online database service
- explain web-based information services and online searching
- describe how the search engine work.
- explain data mining and data warehousing and library expert systems.

5.1 INTRODUCTION

A database consists of an organized collection of data for one or more uses, typically in digital form. One way of classifying databases involves the type of their contents, for example: bibliographic, document-text, statistical. Digital databases are managed using database management systems, which store database contents, allowing data creation and maintenance, and search and other access.

Databases enable you to identify, locate and obtain information resources across a wide range of disciplines, and to search on very specific and

NOTES

detailed topics. For example, databases can be used to find journal articles and chapters in books. Becoming proficient in using electronic databases equips you with a powerful and indispensable research skill.

5.2 DATABASE SEARCH SERVICES

Libraries all over the world are finding vital information through online information retrieval services that provide access to thousands of databases containing both bibliographic and primary source information. Databases are also available as discrete datasets on optical discs like CD-ROM, DVD-ROM, etc. Users can either search these databases directly or through intermediaries (such as library professionals). Databases that are available in libraries for remote access via online search or for local access via CD-ROM/DVD-ROM can be categorised as:

- *Reference databases:* They refer users to another source such as a document, an organisation, an individual or full text of a document. These may be grouped as:
  - *Bibliographic databases:* Provide information on contents, location and summarisation through citations, bibliographic references and abstracts.
  - *Catalogue databases:* Provide information on the stock of a given library or group of libraries. These databases generally include monographs, journal titles and other library materials.
  - *Referral databases:* Include directory type data such as the names and address of organisations or individuals.
- *Source databases:* Contain original source data and act as one type of electronic document. These are grouped as:
  - *Numeric databases:* Contain numerical data *e.g.*, statistics and survey data.
  - *Full text databases:* Contain databases of newspaper items, journal articles, patents etc.
  - *Alphanumeric databases:* Contain textual as well as numeric data *e.g.*, annual reports, handbooks etc.
  - *Multimedia databases:* Contain information stored in a mixture of formats *e.g.*, text, sound, video, picture animation etc.

All the above mentioned databases are accessible remotely through online services or locally through CD-ROM databases. The search framework of online and CD-ROM databases is quite similar and includes following components:

- *Formulation:* It involves several decisions regarding sources, fields, what to search for, and the search variants. Users need to select first the database(s) to be searched. A search may be conducted against one or more selected fields in a database but a search on specific fields produces precise search result

than one on a complete record. The next task is to write actual search statement. In search statement search terms may be represented in various ways (singular/plural forms, synonymous forms, variant spelling etc.) and combined by using various search operators (Boolean operators, positional operators, relational operators, etc.). The general search techniques are:

- *Boolean search:* It combines search terms according to Boolean logic and allows three types of search – AND, OR, and NOT.
- *Truncation:* It enables a search to be conducted for all the different forms of a word having the same common root. The search term may be represented through right-hand truncation or left-hand truncation. For example, the right truncated term catalog will retrieve records containing catalogue, catalogues, cataloguing, catalog and so on.
- *Proximity search:* It allows users to specify the distance (in terms of character) between two search terms in the retrieved result sets. In some retrieval system it is represented as NEAR operator.
- *Field-level search:* It restricts a search to a specific field with a view to obtain more precise results.
- *Limiting search:* It limits a given search by using certain criteria, such as language, year of publication, type of information sources and so on.

- *Action:* In the second step, a search button needs to be pressed to conduct a search and the user is expected to wait till the search process ends.
- *Review of result:* In this step, user views the retrieved results by selecting various display options such as size of display, display format, order of items (by author, title, date, etc.) and so on.

Refinement: If user is not satisfied with the search result, he/she needs to reformulate the search statement and conduct a new search. Users can also refine a search and conduct a new search on the retrieved set.

5.3 ONLINE DATABASE SERVICE

During the past 10 to 15 years, several related data processing and telecommunication technologies have evolved and merged to make fast, reliable and low cost online information services a reality. The components of any online database service include:

- Powerful time-sharing computers
- Interactive retrieval programs

NOTES

- Telecommunication support
- Access privilege
- Machine-readable databases
- Fast, low cost terminals
- Fast access disc storage devices
- Networks

Although online services have a place in almost any kind of library, they are most commonly found in libraries oriented to scientific and technical research or business. These services provide facilities for retrospective search, comprehensive search, selective search, state of the art search, SDI search, ready reference search and patent search. The advantages of online searching are:

- *Speed:* Online retrieval is much faster than manual searching.
- *Comprehensiveness:* Provides access to many more information sources than even the largest libraries can support.
- *Currency:* Sources are updated before their published counterparts printed and distributed.
- *Flexibility:* Supports interactive retrieval and permits many more access points than manual searching allows.
- *Convenience:* Allows retrieval from anywhere.
- *Cost effectiveness:* An in-depth online search can be conducted at lower cost in comparison to an equivalent manual search of printed resources.

CD-ROM Database Service

Optical discs, particularly in the form of CD-ROMs have become important medium for storage and retrieval of information. CD-ROM databases act as alternative information access system to online database services via telecommunication network and www. CD-ROM databases can be accessed in a standalone PC (single user – single CD-ROM) or over CD-ROM network (multiple user – multiple databases). Multiple access to CD-ROM databases is provided through network file server or dedicated CD-ROM server or jukebox system. CD-ROM databases contain bibliographic datasets, catalogues, source databases, reference databases or multimedia databases. The selection of CD-ROM databases should be based on some well defined criteria, including but not limited to database contents, currency, availability of back files, quality of retrieval software, user interface, printing and downloading facilities, SDI service facility, data access time, cost and standardisation. The process of information retrieval in online services or for CD-ROM databases is quite similar. The common steps are Selection of database, Formulation and entering of search query, Selection of documents from retrieved result set, View/display of selected documents and Print or download in desired format (generally as text file).

NOTES

5.4 WEB-BASED INFORMATION SERVICES

The Internet is a global collection of interlinked computer networks, or a network of networks. It offers a gateway to myriad online databases, library catalogues and collections, software and document archives, in addition to frequently used store-and-forward services such as usenet news and e-mail. The resources available in the Internet can be accessed by a number of services such as Telnet, FTP, Gopher and World Wide Web (WWW). The WWW or simply web is the most innovative, the most visible, and the fastest growing part of the Internet. Web supports multimedia, hyperlinking and HTML formatted web pages can invoke programmes (*i.e.*, CGI) to process user supplied data (*i.e.*, Form). You already have an idea about the integration of web and library automation package in previous units of this block. Web can be utilised in library services as global publishing platform in two ways. We can link local library resources to the web for global users and we can organise global information resources available in the web for the local users. The web-based information services may be discussed under two broad groups – general web-based information services and subject specific web-based information services.

General Services

Web includes a vast array of information resources, including but not limited to:

- *Listserv and Discussion Groups:* Includes a wide variety of topics and provides opportunity to exchange current information and conduct a dialogue;
- *Community Information:* These are local data related to weather, demography, tourist places, historical places, local events, transport etc.;
- *Government Information:* Central, state and local self-governments are providing variety of information on web;
- *Library Catalogues:* Libraries are making their catalogues available over the Internet from all over the world;
- *Commercial Resources:* Financial and commercial data are available from stock exchanges, brokers and other firms;
- *Bulletin Board:* These are electronic newsletters that provide news and factual information;
- *Patent Information:* Almost all the countries are making list of approved patents available through web site;
- *Electronic Journals:* Many primary and secondary journals (indexing and abstracting journals) are accessible through web. Some of these e-journals allow free full-text access. Directories of e-journals on various subjects are also available on the web such

NOTES

as PubList (http://publist.com/search.html) and BUBL (http://bubl.ac.uk/journal);

- *Electronic Books and Book Reviews:* Electronic books appeared first in CD-ROMs (such as reference books) and now these are available through web. Many electronic books are produced that can be used only with specific readers such as RCA e-Book reader, Adobe e-Book reader, Gemstar e-Books reader etc. Some of the significant e-book providers are net Library of OCLC (http://www.netlibrary.com), Questia (http://questia.com) and ebrary (http://www.ebrary.com);
- *Theses and Dissertations:* Theses and dissertations were unpublished sources of information until recently. These primary information sources are now available in the web sites of universities for consultation and downloading; and
- *Education and Training Materials:* Course materials, interactive tutorials and multimedia presentations ranging from school grade to postgraduate grade are available in web for online and offline learning.

Web is a treasure of information. These information resources are either accessible directly through URL (Uniform Resource Locator) or by resource discovery through web searching tools. The web searching tools are categorised into three groups – subject directory, search engine and meta search engine. Libraries generally offer directory-based access to evaluated and useful information resources and web searching tools. These directories are often integrated with OPAC to provide single access interface to users.

Subject-Specific Services

Subject-specific information resources are mainly available in the web through three channels – Academic subject directories, Subject gateways and Digital libraries.

Academic Subject Directories: Web directories organise digital information sources by using metadata schemas, bibliographic classification schemes or subject indexing tools. Some well-known subject directories are:

Cyber Dewey (http://www.anthus.com/Cyber Dewey/CyberDewey.html): It uses DDC in organising digital information resources. Selection of a Dewey class takes users to the specific subdivision of the class with items listed against each subclass.

The Information Search Process

Basic Features

Whether it is manual or online searching, basic features of search processes are same. These are:

(*a*) matching of information needs with the information available in IR system;

(*b*) conducting the search with some criteria such as surrogates, exact form, etc.; and

(*c*) subject surrogates are the main approach of information seekers.

Search Tactics

The success of any search depends upon the tactics effectively adopted in conducting a search. Search tactics is concerned with adopting techniques in manoeuvring the actual search. The tactics described by Bates can be grouped in four categories:

(*a*) Monitoring tactics (to keep the search on a track)

(*b*) File structure tactics (for traversing information within the system)

(*c*) Search formulation tactics (to aid in the process of designing and redesigning the search formulation)

(*d*) Term tactics (to select and reuse search terms)

In order to conduct a successful online search one must perform the following tasks:

(*a*) Decide which particular database(s) is/are to be searched;

(*b*) Guess the words that might have been used by the authors and the indexers in a database of potentially relevant documents;

(*c*) Use the thesaurus of the chosen database in order to translate the query terms in the appropriate way (is in the language of the system);

(*d*) Coordinate the terms (often using Boolean operators) to formulate the search statement;

(*e*) Use the search features and search operators appropriate for the chosen database;

(*f*) Input the search statement;

(*g*) Repeat steps (*e*) and (*f*) until a desirable output is obtained or the search fails altogether; and

(*h*) Identify the actual relevant items from among those retrieved.

One major task in the searching process relates to the coordination of terms [step (*e*) above] in order to formulate the actual search statement. The result of the search depends largely on how adequately the search terms are combined. Boolean search techniques have been used widely since the beginning of mechanised information retrieval. However, other models, especially the vector-space model, or a variation of it is also used in modern information retrieval systems. In any case, the basic information search process that a user has to follow remains the same for almost all online information retrieval systems.

NOTES

In the following sections we shall see the major steps to be followed while conducting an information search using one of the oldest and most widely used online search services, viz. the Dialog information services.

Multiple Database Searching and Common Command Language

Databases are generated by various agencies in different subject fields and database producers use different commands and instructions for searching the databases. Different software packages also provide different command languages. As a result users need to be conversant with different command languages and have to have sufficient knowledge if they intend to search more than one database. As it is known, thousands of databases are now existing or are being developed. Thus, it becomes difficult for any user to develop mastery in all of them, even if he has to search small number of databases in his own subject/discipline. Some online database vendors like Dialog, STN, who provide access to several databases, have solved the problem to some extent by adopting some command language to search the databases serviced by them.

To solve the problem of heterogeneity in command languages, some efforts are being made at international level. Z39G Committee of National Information Standards Organisation (NISO), USA, set up in 1980 came up with a draft Common Command Language (CCL) in 1987. The International Organization for Standardization (ISO) also came up with a standard CCL in 1988 (ISO 8777). There is great similarities between these two standards. However, implementation of these is still awaited. Wide implementation of the CCL will facilitate the users in reaching multiple databases of different producers and vendors with less effort. With this, for database producers and users it will be advantageous to pay more attention to the contents part while searching. It will also facilitate the trainers to pay more attention to the intellectual part of content of databases rather than spending more time on the technicalities of searching the databases.

5.5 ONLINE SEARCHING

Online information retrieval involves searching remotely located databases through interactive communication with the help of computers and communication channels [*Chowdhury, 2004*]. The database can be accessed by the user directly or through a vendor (supplier of online services); in each case through the computer and communication network. The term online retrieval can thus be used to indicate the information retrieval services available from producers of databases, or vendors of these databases. Although online information retrieval systems

NOTES

have existed for more than four decades, emergence of the Internet and world wide web have brought significant changes and improvements in the online information retrieval environment.

The phrase online searching was originally used to describe the process of directly interrogating computer systems to resolve particular requests for information. Now the phrase is used to denote searches that are conducted by means of a local computer that communicates with a remote computer system containing databases. Users can access the database(s) via an online search service provider (also called vendor). The search process is interactive and the user can conduct the search iteratively until a satisfactory result is obtained. With the advent of the Internet and world wide web, the connotation of online searching has changed. Now we can conduct online searches through the world wide web on information sources that are distributed all over the world. For searching these information sources through the web, we can go straight to the web page of the service provider, provided we know the URL (uniform resource locator, or the address of the web page). Alternatively, we can try to locate the information source(s) by searching through the web search engines (the retrieval programs that help us search the web) like AltaVista and InfoSeek, or through subject directories or gateways (these are hierarchically organized lists that can be navigated to reach a particular information source or a group of similar sources) like Yahoo, SOSIG and Biz/ed.

In the following sections we shall first discuss the former type of online service – the traditional online service characterised by a remote online database search service offered commercially by a search service provider or vendor. We shall look at the online information search process with Dialog Information Service as an example. We shall then discuss the information retrieval from the world wide web with a view to understanding how traditional online searching differs from web information retrieval services.

One major advantage of the traditional online search services is that they are designed to be pay-as-you-go, and therefore each search session can be costed. Another advantage of online searching is its speed and the currency of the data retrieved. Originally online search services were very expensive and could be complex, and therefore intermediaries were needed to help end-users conduct an effective and efficient online search. However, over the years online search services have become less expensive and more user-friendly. As a result, they can now be used by end-users themselves, and can be accesses through the web.

Online Search Services

There are various components of an online search service such as:

- information providersor database producers who provide databases to be accessed in an online mode;

NOTES

- a search service provider or vendor, which provides access to the databases and software for conducting the search;
- communication links that connect the user with the host and the database(s); nowadays users can communicate with the service providers through the web and therefore an Internet connection is necessary; and
- a local workstation through which the user is linked to the service.

Online search services, or vendors, are those organisations that provide value-added processing to the databases and offer search services. The following are some examples of online search services:

- Dialog (http://www.dialog.com/about/): A pioneer in online search services, Dialog provides online access to over 800 million records in 900 databases in different disciplines.
- OCLC FirstSearch (http://www.oclc.org/firstsearch/): This provides library users with instant online access to more than 72 databases, including these valuable OCLC databases: OCLC WorldCat, OCLC FirstSearch Electronic Collections Online, OCLC ArticleFirst, OCLC PAIS International, OCLC PapersFirst, OCLC Proceedings First and OCLC Union Lists of Periodicals.
- Ovid (http://www.ovid.com/site/index.jsp): Ovid provides access to hundreds of full text journals, renowned textbooks and premier bibliographic databases in various disciplines.
- STN (http://www.cas.org/stnonline.html): STN offers current and archival information from over 200 scientific, technical, business and patent databases covering a broad range of scientific fields, including chemistry, engineering, life sciences, pharmaceutical sciences, biotechnology, regulatory compliance, patents and business.

While the above are examples of online search services that provide online access to a large number of databases in various disciplines, the following are examples of some online search services that provide access to the full texts of journals and books:

- EBSCO Information service (http://www.epnet.com/default.asp): This provides access to a large collection of full text and bibliographic databases suitable for all kinds of libraries.
- Ingenta (http://www.ingenta.com/): Ingenta provides access to the full text of over 5,400 publications from over 230 academic and professional publishers.
- ProQuest (http://proquest.umi.com/pqdweb): ProQuest is a resource of electronic collections containing millions of articles originally published in magazines, newspapers and journals.

Details of online search services are available in a number of publications [see for example, Forrester and Rowlands, 1999; Large, Tedd and Hartley, 1999; and Chowdhury and Chowdhury, 2001b] as well as in the websites of the respective search services.

Basic Steps in an Online Search

The steps involved in carrying out an online search vary from system to system. This is because each system has its own custom-built interface, which allows specific types of search and uses specific operators for different search commands. Nevertheless, the graphical user interfaces used in these systems have made the task of searching reasonably straightforward and the process of searching has been simplified further in the web-based interfaces. These are the basic steps that one needs to follow to conduct an online search:

(*i*) Study the search topic and develop a clear understanding of the information requirement. This is a critical step and depends on a number of factors, such as, the nature and requirements of the user, how well the user can express his or her information needs, how much the user already knows, how the user is going to use the information, and so on. This happens before the actual search process begins and is often conducted through a series of dialogues between a searcher and an information intermediary. In the absence of an intermediary, users have to clearly delineate their information requirements for themselves. This stage decides the strategies to be adopted for a search.

(*ii*) Get access to an online search service. This can be done through subscription or a licensing agreement. The access right has to be obtained before the search begins.

(*iii*) Log on to the service provider. Nowadays this is usually done though the web interfaces of the online search service providers. Users need to know the URL of the online service provider as well as the user login-ID and password. Sometimes Internet Protocol (IP) authentication mechanism is adapted instead of user login-ID and users directly logon to the portal.

(*iv*) Select the appropriate database(s) to search. This is a critical and often a difficult task. The success of a search largely depends on the appropriate selection of the databases. Online search services allow users to select one or more databases to search using the same interface. Most search services allow users to browse through the database categories to select appropriate databases(s). Dialog has a unique facility called DialIndex search (details are given later in this unit), which allows users to see how many times a given search term occurs in a set of chosen databases. This information can guide users to select the appropriate database to conduct the actual search.

(*v*) Formulate search expressions. This is the key part of the job. It may involve a number of activities, the first being the selection of appropriate terms and/or phrases. This may require the user to consult dictionaries and thesauri. Once the appropriate search terms and/or phrases are chosen, the search expression has to be formulated. At this stage the user should have an understanding of the nature, content and structure of the chosen database(s) and to know which fields are indexed and therefore can be searched. The user also needs to know what search facilities are available, such as Boolean search, truncation, field specific search, proximity search, and so on, and the appropriate operators. The search operators and syntax for formulating search expressions vary from one search service to the other. Many search service providers have different interfaces for novice and expert users. If the users want to use the expert search interface, which may be command-driven, they have to have a knowledge of the various search commands and their order of execution.

(*vi*) Select the appropriate format for display. Online search services allow users to select an appropriate format, from a number of predefined formats, to display the retrieved records. However, there may be charges for the records displayed. For example, when searching Dialog, charges incurred include output and search time costs, as well as Internet charges; prices also vary by database. Therefore, one has to be very careful in deciding which record(s) display and in which format to display them. If the option for the display of the full record(s) is chosen, the process may take some time, depending on the network traffic. However, each online search service provides an option for brief display, which shows the brief details of the output records, and users may select records from this list for a full display.

(*vii*) Reformulate your query, if necessary. This may mean going back to step 4 or step 5 and repeating the entire process. Online searches are usually iterative processes, meaning that user conducts several searches, compares the results, modifies a search statement, or conducts a new search in order to get the best results.

(*viii*) Select the mode of delivery. You may download all the chosen records online or send an offline request.

In the following section, the methodology of DialogWeb search is provided to give you an idea about how online search can be conducted.

Features of an Online Search Service: DialogWeb

DialogWeb is the web interface to the Dialog online search service, one of the oldest and largest online search service providers, which gives easy access to a large number of databases with:

- company information–both directory listings and financial information
- industry information–trends; overviews; market research; specialised industry newsletters and reports; U.S. and international news, including an extensive collection of newspapers and newswires from North America and Asia; and U.S. government news, including public affairs, law and regulatory information
- patents and trademarks – a worldwide collection for research and competitive intelligence tracking
- chemistry, environment, science and technology – technical literature and reference material to support research needs
- social science and humanities including education, information science, psychology, sociology and science, from public opinion, news, and leading scholarly and popular publications
- general reference information – people, books, consumer news and travel.

Users can search and retrieve information from all these different types of information sources using:

- guided search mode, which does not require knowledge of the Dialog command language;
- command search mode, which allows experienced users to use the Dialog command language;
- database selection tools, which help users pinpoint the right database for a search;
- integrated database descriptions, pricing information and other search assistance; and
- easy to use forms to create and modify Alerts (current awareness updates).

Dialog search results are available in HTML or text formats. Users have a choice of displaying records or sending search results via email, fax, or postal delivery.

Steps in a DialogWeb Search

The first step of a DialogWeb search involves logging in to the system, for which a Dialog account is necessary. The user goes to the DialogWeb site (http://www. DialogWeb.com) and must enter the user ID and password. The log-in screen also provides information about DialogWeb and a preview and search tips. After logging in, the user needs to select the mode of search: Guided Search or Command Search. Guided Search is the default search option.

Guided Search

Guided Search is designed for novice to intermediate searchers. The following steps are to be followed for conducting a Guided Search looking for information on the topic of 'digital libraries'.

NOTES

Step 1: Choose Database

To begin the Guided Search, the user clicks the New Search button and chooses from the list of main subject categories. Each category is further divided into focused search topics. For a search on digital libraries, one can select these categories:

Social Sciences and Humanities > Social Sciences > Library and Information Science.

This will lead to a list of databases that cover Library and Information Science.

Step 2: Choose a Search Option and Carry Out Search.

In Guided Search there are two search options:

- Targeted Search, which is available in some, but not all, subject categories. It is a ready-made search form with databases pre-assigned to the form.
- Dynamic Search, which is available in all the subject categories. The Dynamic Search form is generated based on the category or database that is selected.

  Dynamic Search has access to many more databases compared than the Targeted Search and is more flexible.

Targeted Search is the easiest type of search to perform. The user can enter the search word or phrase as 'Words in Title' or as the 'Main Subject'.

Dynamic Search is available at various points in the search category selection process or when a user chooses the Quick Functions option in New Search and enters a specific database number. The Dynamic Search capability is available no matter what category or database is picked. In a category with many databases assigned to it, a user can search:

- all of the databases together
- a group of similarly designed databases together
- one of the assigned databases individually.

If a user has chosen the Dynamic Search option and has decided to conduct the search on all the databases under the 'Library and Information Science' category, the 'Dynamic Search' screen is shown. The Dynamic Search forms also offer the following options:

- Navigation – The search category selections display at the top of the form. To return to a category or option, the user clicks the search category or option name.
- Run Saved Strategy – If a user has already saved a search strategy, it can be run against the selected databases by clicking Run Saved Strategy.

A list of the databases used in the search is displayed at the bottom of the form.

The info (*i*) icon gives more information about the database content and pricing. In the Dynamic Search screen users can enter a search term or phrase and conduct the search on subject, author, descriptor or title field. A search can also be restricted by the year of publication, and the user can browse the list of items by author or year of publication.

Step 3: Display Search Results

The search results from a Targeted Search or a Dynamic Search will appear on a Picklist page, which provides a quick view of the records. From the Picklist page users can choose to:

- display specific records in more detailed formats or send records via e-mail or fax, or by post
- rearrange the order in which the records are displayed
- refine the search strategy
- remove duplicate records
- view the prices for all format options
- save the strategy for future use
- create an Alert for automatic updates on the search topic.

After the search has finished processing, the Picklist page will appear. Users can choose to view results by selecting one or more items by checking the boxes and then selecting the display button, or can display any one record just by clicking on the hyperlinked title. The format for display is chosen from the 'Format' box and the records are sorted according to a sort criterion chosen from the 'Sort by' box. The search expression can be refined by clicking the 'Back to Search' button, which allows users to edit, add or delete information from the search form.

Command Search

Command Search is designed for intermediate to experienced Dialog searchers. It provides complete command-based access to Dialog's extensive collection of databases. Users are expected to be familiar with the various Dialog commands when using Command Search. Additional features include built-in tools such as Bluesheets (database descriptions) and pricing information, database selection assistance to help pinpoint the right databases for a search and easy to use forms to create and modify Alerts (current awareness updates). The Command Search main page allows users to begin inputting Dialog commands immediately. A Command Search contains:

- a textbox for entering Dialog search commands
- a Submit button that sends the command

NOTES

- a Previous button that displays your most recent command entries.

The main page has links to the Databases feature, product support information, and Guided Search. Users can move between Guided Search and Command Search. Steps for conducting a Command search can be summarised as follows.

Step 1: Choose Database(s)

DialogWeb simplifies database selection by arranging the databases by subject in the Databases feature. Users can select one or more databases by checking in the Database box. However, if users are not sure which database(s) to select, they can choose the Dialog Index option. This is particularly useful when users do not know which databases to search, or when they want to carry out a comprehensive search and cover everything on a topic. Dialog Index is a master index to most of the Dialog databases, and it allows users to compare the number of records retrieved from a group of databases.

After selecting a database, users must search the databases to view the records. They can click 'Begin Databases' to enter the files that they have checked and run the same strategy, or may choose the database(s) to search by entering the file numbers and even change their search strategy in the command line.

Step 2: Choose a Search Option and Carry Out Search

Once the databases are chosen, the Dialog Command Search page appears. It can also appear:

- after log-in if it is set as the default
- when the Command Search link from the main Guided Search page is clicked
- when the Begin Databases button from Databases is clicked for browsing.

The appropriate BEGIN or 'b' command is inserted in the command line automatically when a search has been made in Databases and one or more databases have been selected. Users can add the CURRENT command to their BEGIN statement by typing in 'current' after the command. This allows them to search the current year and one year earlier, and narrows the search results at the beginning. Then they click the Submit button or press the ENTER key on the keyboard to start the search.

Any Dialog command followed by the search term(s) should be entered in the search box. The terms when looking for information on digital libraries might be: S digital (w) libraries.

Step 3: Add Operator to a Search

The search can be refined by including 'electronic libraries' through the following expression:

S (digital or electronic) (w) libraries

This search statement will retrieve records on electronic as well as digital libraries. The search statement retrieves those records where digital and libraries, or electronic and libraries, occur next to each other in the same sequence. More records are retrieved by truncating the search term libraries as follows:

S (digital or electronic) (w) library?

Various other modifications may be made by using appropriate search commands, for example, limiting the search to one or more fields or limiting the results to a language, year of publication, and so on.

Step 4: Displaying Records

A search history of all of the sets appears and users can view some of the records. It is a good idea to display a few records in 'free' format before displaying the records in the long or full format. To display records users can choose a format from the drop-down list and click Display for the appropriate set. Formats determine the amount of information to be displayed for each record. The Format list box lists the basic format options: free, short, medium, long, full and KWIC. It is possible to indicate the number of records to display; the default is 10 and a maximum of 99 records can be specified. There is an option of using a Type command to display records or From Each together with the Type command, in order to search more than one database.

Access to Information on the Web: The Tools

A user can get access to any website by entering the URL (Uniform Resource Locator; the address of a web site) on the browser. A web browser, like Netscape Navigator or Microsoft Internet Explorer, is a computer program, an essential tool for getting access to the web. A web browser performs two major tasks:

- It knows how to go to a web server on the Internet and request a page so that the browser can pull the page through the network and into your machine.
- It knows how to interpret the set of HTML (Hypertext Markup Language; a language format used to create web pages; HTML is called the lingua franca of the web) tags within the page in order to display the page on your screen as the page's creator intended it to be viewed.

Although one could get access to any web page by typing on the browser the URL of a sought website, and then moving into the site through the various links, there are several problems to this approach, especially when the user is interested to get some specific information on a given topic, or find answers to a given question. Problems arise because it is almost impossible for users to know which of the billions

of web pages may contain the information they require and which of the millions of websites contains the required web page. In order to solve these difficulties, several web search tools have been developed that assist users in finding the information they need from the right web page with relatively little effort.

There are basically two ways to find information on the web: by conducting a search using what is known as a search engine, or by following the links in a specially designed list called a directory. Search engines allow users to enter search terms–keywords and/or phrases–that are run against a database containing information on web pages collected automatically by programs called spiders. The search engine retrieves from its database web pages that match the search terms entered by the searcher. It is important to note that when a user conducts a search using a search engine, the search engine does not search for the information across the entire web at the given instance. Instead, it searches a fixed database, which is updated at a regular intervals according to a specific set of criteria employed by the search engine, located at the search engine's website and containing information on selected web pages.

5.6 HOW THE SEARCH ENGINES WORK?

Although all search engines are intended to perform the same task, each goes about doing so in a different way, sometimes with very different results. Factors that influence the search results include the size of the database, frequency of updating it, criteria employed for indexing the chosen web pages, and the search engine's retrieval capabilities. Search engines also differ in their search speed, the design of the search interface, the way in which they display the search results, the amount of help they provide, and in other ways.

Search engines run from special sites on the web and are designed to help people find information stored on other sites. There are differences in the ways search engines work, but they all perform the following three basic tasks:

- They search the Internet–or select parts of the Internet–according to a set of criteria.
- They keep an index of the words or phrases they find, with specific information such as where they found them, how many times they found them, and so on.
- They allow users to search for words or phrases or combinations of words or phrases found in that index.

There are three main components of a search engine: the spider, the search engine software and interface, and the index.

NOTES

The Spider

To find information from the millions of web pages, a search engine employs special software called a spider or crawler. It is a program that automatically fetches web pages for search engines; it is called a spider because it crawls over the web. Spider programs treat the web as a graph and, using a set of URLs as a seed set, traverse the graph to select web pages. The crawler traverses the graph either breadth first (searching all nodes at one level of the tree before going down a level) or depth first (searching the current path as far as possible before backtracking to the last choice point and trying the next alternative path in the tree). Web pages contain links to other pages, and a spider uses these links to move to another page; it visits the page, reads it and then follows links to other pages within the site.

One of the major problems for a spider or a crawler program, and indeed for a specific search engine, is to decide which page to select for indexing. Each search engine aims to index the most important web pages, and therefore aims to prioritize URLs to obtain the best pages. The quality of a web page may be judged in many ways, for instance by measuring its content, by assessing its popularity (by counting the number of visits) or by measuring its connectivity (which other pages link to this page). Spider program metrics based on connectivity have the advantage that they do not require information that is not easily accessible (such as page popularity data), and that they are easy to compute, so they scale well to even very large page collections. Another major issue for a crawler program is to schedule the frequency of revisiting pages. Since web pages keep changing constantly it is important to visit them frequently. Search engines do not usually disclose the details of their spider programs.

Search Engine Software

Search engine software is the information retrieval program that performs two major tasks: it searches through the millions of terms recorded in the index to find matches to a search and it ranks the retrieved records (web pages) in the order it believes is the most relevant. The criteria for selection (or rejection) of search terms and assigning weight to them depend on the policy of the search engine. Similarly, the specific information that is stored along with each keyword – such as wherein a given web page it occurred (heading, links, meta-tags or title of the page), how many times it occurred, the attached weight, and so on – depend on the policy of the search engine concerned [Sullivan, 2003]. Each commercial search engine has a different formula for assigning weight to the words in its index. This is one of the reasons for the fact that a search for the same word on different search engines may produce different results, and the retrieved web pages may be ordered differently.

Google uses the concept 'page rank' to determine the importance of a web page [Brin and Page]. The idea is based on the principle of citation analysis. A document, according to the basic principle of citation analysis, is considered to be important if it is cited frequently, and thus one can rank a set of documents in order of their importance by counting the number of times each one has been cited. A web page's 'page rank' is an objective measure of its citation importance that corresponds well with people's subjective idea of importance.

Indexing

Early search engines indexed only components of each web page, but increasingly full texts of web pages are indexed [*Rasmussen, 2002*]. Specific information on the form and weight of index terms, the techniques for calculating relevance, and so on are usually proprietary. Most search engines use variations of the Boolean and vector space model. Although search engines do not usually disclose their secret of ranking pages, some general information is available. One of the main rules in a ranking algorithm involves the location and frequency of keywords on a web page [Sullivan, 2003]. The location of a term on a page is also used as an important criterion. Pages with the search terms appearing in the HTML title tag are often assumed to be more relevant to the topic than others. Pages where the search terms appear near the top of a web page, such as in the headline or in the first few paragraphs of text, are also ranked highly compared with others. Frequency of occurrence is the other major factor used to determine relevance. Pages where the search terms occur frequently are often deemed more relevant than other web pages.

Features of each search engine can be learnt by following the help pages. However, there are a number of sources that regularly report on the features and comparisons of web search engines. The most prominent are SearchEngineWatch.com and the Online journal.

Search Strategy

A search is performed to obtain exact or related information about a topic/keyword/subject area etc., from a repository of information. Search strategy is the systematic tactics to be followed in finding the relevant information on a topic. Thus, a proper strategy is vital to obtain the best search results. The best search results can be described as the information that matches as closely to the information that is required by a search formulator. Therefore, good search results are not only based on the well organised information, but also on the use of right search formulation. In other words a good search strategy may lead to good search results using minimal efforts and cost.

A search in a database may be a combination of an interactive, iterative and heuristic process. Often it is a trial and error exercise. A searcher

interacts with a system for a search. However, a single formulation of search may not bring in best results, thus, the search may require interaction with the end user. Every iteraction refines the search on the basis of information results of previous iteraction.

NOTES

Information may be searched by the end-user directly, or by an information professional – an intermediary *e.g.*, reference librarian. Success of a search depends on the end-user and his/her effective communication skill either directly with the system or with the intermediary. Let us look into some stages of searching in the next section.

The stages in database search may be summarised as follows:

1. Recognition of an information need: Defining the need for information by the end-user in terms of a specific subject, type of information/ document desired. For example, one can specify a bibliographic reference with or without abstract, or how soon the information is needed and other details.
2. Communication of the information need to the information/ database service center: This step is specifically needed when search is not to be done by an end user himself/herself. This step can be done in person, by letter, telephone, e-mail, or through another person.
3. Recording of the search request: Search request with details is recorded in a Search Record Form (online or printed).
4. Specification of search request: If necessary and if possible, the reference librarian/information specialist should arrange for a discussion with the end-user and specify the information need as precisely as possible. Aids, such as, a scheme of classification or a thesaurus covering the subject of the query, subject map, known documents or a specialist in the subject of the query can help in the understanding of the topic. The display of these aids can be done online in many cases for the end-user to view.
5. Select the database resource: Selection of appropriate database(s) likely to yield best results.
6. Formulate query in search language: Formulation of the query is to be made in the search language of the database and/or the software used in the creation of the database. This implies that the adoption of search strategy and search expressions should be appropriate to the structure, organisation, search language and capabilities of the system. Vocabulary management tools, such as, thesaurus, classification scheme, subject heading lists, etc. associated with the database(s) to be searched assist this step.
7. Perform the search operation: Perform search on fast access files, involving indexes yields faster result. The end-user can assist in the evaluation of the initial/intermediate search results and provide feedback.

NOTES

8. Modification of search: If necessary, modification of search strategy and refinement of the search expression can be done on the basis of the successive end-user feedbacks on the intermediate retrieval results.
9. Displaying information as per user preference: Selecting the most relevant references/abstracts and arranging the result as desired by end-user.
10. Recording/logging the query: The search procedure adopted and the results obtained can be recorded for future use and analytical studies.

5.7 DATA MINING AND DATA WAREHOUSING

A Data Warehouse is a:

- Subject-oriented
- Integrated
- Time-Variant
- Nonvolatile

Collection of Data is Support of Management's Decision

The defining characteristics of a data warehouse are:

- **Subject-orientation:** Data warehouse data are arranged and optimized to provide answers to questions coming from diverse functional area within a company. Therefore, the data warehouse contains data organized and summarized by topic, such as sales, marketing, finance, distribution, and transportation. For each one of these topics the data warehouse contains specific subjects of interest-products, customers, deparatments, regions, promotions, and so on. Note that this form of data organization is quite different from the more functional or process-oriented organization of typical transaction systems.
- **Time-variancy:** We have already noted that the DSS data include a time element. In contrast to the operational data, which focus on current transactions, the warehouse data represent the flow of data through time. The data warehouse can even contain projected data generated through statistical and other models.
- **Non-volatility:** Once data enter the data warehouse they are never removed. Because the data in the data warehouse represent the company's entire history, the operational data representing the near-tern history, are always added to it. Because data are never deleted and new data are always added, the data warehouse is always growing. That is why the DSS DBMS must be able to support multi-gigabyte and even multi-terabyte database and multiporocessor hardware.

- **Integration:** The data warehouse is a centralized, consolidated database that integrates data derived from the entire organization. Thus the data warehouse consolidates data from multiple and diverse sources with diverse formats. Data integration implies a well-organized effort to define and standardize all data elements. This integration effort can be time-consuming but, once accomplished, it provides a unified view of the overall organizational situation. Data integration enhances decision-making and helps managers to better understand the company's operations. This understanding can be translated into recognition of strategic business opportunities.

Data Mining

Why is data mining being put to use in more and more businesses? Here are some basic reasons:

- In today's world, an organization generates more information in a week than most people can read in a lifetime. It is humanly impossible to study, decipher, and interpret all that data to find useful patterns.
- A data warehouse pools all the data after proper transformation and cleaning into well-organized data structures. Nevertheless, the sheer volume of data makes it impossible for anyone to use analysis and query tools to discern useful patterns.

  In recent times, many data mining tools suitable for a wide range of applications have appeared in the market. The tools and products are now mature enough for business use.
- Data mining needs substantial computing power. Parallel hardware, databases, and other powerful components are available and are becoming very affordable.
- Organizations are placing enormous emphasis on building sound customer relationships, and for good reasons. Companies want to know how they can sell more to existing customers. Organizations are interested in determining which of their customers will prove to be of long-term value of them. Companies need to discover any existing natural classifications among their customers so that the each such class may be properly targeted with products and services. Data mining enables companies to find answers and discover patterns in their customer data.
- Finally, competitive consideration weigh heavily on organizations to get into data mining. Perhaps competitors are already using data mining.

Data Mining Techniques

Data mining covers a broad range of techniques. Each technique has been heavily researched in recent years, and several mature and efficient algorithms have evolved for each of them. The main techniques are: *Cluster detection, Decision trees, Memory based reasoning, Link analysis, Rule induction, Association rule discovery, Outlier detection and analysis, Neural networks, Genetic algorithms,* and *Sequential pattern discovery.* Discussion on the algorithms associated with the various techniques has been kept outside the scope of this text for two main reasons: firstly, because they are too mathematical/technical in nature, and secondly, because there are numerous, well written text books, to serve the needs of those who are specially interested in the subject. For example, a decision tree model may actually be implemented through SQL statements. In the framework, the basic process is the process performed by the particular data mining technique. For example, the decision trees perform the process of splitting at decision points. How a technique validate the model is important. In the case of neural networks, the technique does not contain a validation method to determine termination. The model calls for processing the input records through the different layers of nodes and terminate the discovery at the output node.

5.8 LIBRARY EXPERT SYSTEMS

Expert systems are computer programs that are derived from a branch of computer science research called Artificial Intelligence (AI). AI's scientific goal is to understand intelligence by building computer programs that exhibit intelligent behaviour. It is concerned with the concepts and methods of symbolic inference, or reasoning, by a computer, and how the knowledge is used to make those inferences will be represented inside the machine. In the following paragraphs let us study the views of experts in understanding the concept of expert systems.

Specialised computer programs, modeled on the way human experts tackle problems and arrive at solutions are referred to as 'Expert Systems'. Such systems rely upon a store of specialized knowledge for solving problems and hence also referred to as Knowledge Based Computer Systems (KBCS) or Knowledge Based Systems (KBS).

Expert systems are sophisticated computer programs that manipulate knowledge to solve problems efficiently and effectively in a narrow problem area. Knowledge based systems enhance the value of expert knowledge by making it readily and widely accessible. Like human experts, these systems use symbolic logic and heuristics to find solutions.

They are also capable of learning from experience through inferencing mechanism. According to Liebowitz 'the role of experts systems is to better understand how humans think, reason and learn' [*Leibowitz, 1990*].

NOTES

AI programs that achieve expert-level competence in solving problems in task areas through the knowledge about specific tasks are called knowledge-based or expert systems. Often, the term expert systems is reserved for programs whose knowledge base contains the knowledge used by human experts, in contrast to knowledge gathered from textbooks or non-experts. More often than not, the two terms, expert systems (ES) and knowledge-based systems (KBS) are used synonymously. Taken together, they represent the most widespread type of AI application. The area of human intellectual endeavor to be captured in an expert system is called the task domain. Task refers to some goal-oriented, problem-solving activity. Domain refers to the area within which the task is being performed. Typical tasks are diagnosis, planning, scheduling, configuration and design.

Building an expert system is known as knowledge engineering and its practitioners are called knowledge engineers. The knowledge engineer must make sure that the computer has all the knowledge needed to solve a problem. The knowledge engineer must choose one or more forms in which to represent the required knowledge as symbol patterns in the memory of the computer – that is, he (or she) must choose a knowledge representation. He must also ensure that the computer can use the knowledge efficiently by selecting from a handful of reasoning methods.

Expert Systems for Information Processing and Retrieval

A few expert systems in the field of information processing, information analysis and retrieval are discussed below.

COMIT: COMIT is an early AI language developed at MIT (Massachusetts Institute of Technology). The program is designed for information retrieval. It matches the contents of a highly structured query to highly structured descriptions of biographical sources [Weil, 1968]. A categorization scheme is used as a basis for inferencing and the system follows a pattern matching procedure. Reference sources are described in functional rather than bibliographic terms. The hits or successful matches are ranked according to the degree of certainty that they could answer a query. A declarative knowledge representation scheme is employed for COMIT.

REFSEARCH: REFSEARCH was developed by Joseph Meredith and team, and attempted representation of entire universe of reference works. This program is developed as an instructional tool for teaching

NOTES

librarians the basic principles of reference work. The system aims to study the "principles that would apply to the collection as a whole, to the sum of the data contained in the collection, and to networks of paths leading to the data" [Meredith, 1971]. REFSEARCH describes the various reference sources in terms of the functions they performed. The focus is to describe the reference works as specific types of tools likely to resolve types of information problems. The schemata for knowledge representation in REFSEARCH is quite similar to the methods adopted in present expert systems.

RESEDA: RESEDA developed by Zarri, is a system based on restructuring the knowledge encoded in printed texts. The Reseda system is based on the principles of linguistics. The purpose of RESEDA is recording information found in standard printed works on medieval French biography and history. The goal of Reseda is to use the knowledge base derived from the texts to answer natural language queries about the domain [Zarri, 1985].

CANSEARCH: Politt [1986, 1987] has developed Cansearch, an expert system for access to cancer therapy literature in the Medline database. Cansearch is a hybrid system using rules and frames for knowledge representation. Controlled vocabulary terms and hierarchical relationships between concepts expressed in the Medical Subject Headings are used to guide the search process. Queries formulated are based on knowledge of typical searches, concepts within the domain and the Medline query language.

IR-NLI (Information Retrieval–Natural Language Interface): IR-NLI translates the user's initial query into a formal problem statement. It then models the intermediary's behaviour by consulting a knowledge base of expert intermediary knowledge. The knowledge base contains rules relating to tactics, strategies, and approaches used by searchers. Domain knowledge fed is terminological knowledge about the subject domain of the target database. A formalizer module then generates formal, syntactically correct search strategies [*Brajnik and Tasso, 1986*].

Components of Expert Systems

In general, expert systems are composed of basic components such as:

(*a*) a user interface: to facilitate user interaction

(*b*) a knowledge base: the facts or knowledge based upon which the ES makes decisions

(*c*) an inference mechanism: the reasoning engines built according to heuristics reasoning or facts.

The knowledge base of expert systems contains both factual and heuristic knowledge:

Factual knowledge is that knowledge of the task domain that is widely shared, typically found in textbooks or journals, and commonly agreed upon by those knowledgeable in the particular field.

Heuristic knowledge is the less rigorous, more experimental, more judgmental knowledge of performance. In contrast to factual knowledge, heuristic knowledge is largely based upon past experiences and knowledge gained. For instance the knowledge that, if the same pattern of events take place then the same conclusion could be expected. It is the knowledge built from 'thumb rules'.

Intelligent Information Retrieval

Sparck Jones [1983] defines an intelligent information retrieval system as a computer system with inferential capabilities such that it can use prior knowledge to establish a connection between a user's (probably ill-specified) request and a candidate set of relevant documents. According to Brooks [1987] an intelligent information retrieval system is a system that carries out intelligent retrieval. Brooks further defines intelligent retrieval as the use, by a computer system, of the stored knowledge of its world of documents, users, etc., and of information about the user and his/her problem to infer which documents would enable that particular user to resolve or manage his/her problem in a better way. It has become apparent through different experiments that retrieval cannot be carried out intelligently unless the system 'knows' about its task, world of documents, language, subject domains, etc., as well as the specific requirement of the user. The realization of the need to use knowledge within retrieval systems has led researchers to look at the disciplines of artificial intelligence and expert systems that also aim to incorporate and use knowledge.

Expert Systems for Information Processing and Management Applications

Over the years researchers have developed several expert systems for professional tasks in both traditional and non-traditional library and information services and management. These tasks include: indexing, abstracting, thesaurus construction, cataloguing and classification, Boolean text retrieval, non-Boolean text retrieval including reference services, automatic content analysis and knowledge representation, relational database access and management, intelligent documents, training, database selection, and database analysis.

Lancaster and Warner [2001] provide an excellent review of the applications of expert systems and related intelligent technologies in different

areas of library and information science. They note that the major applications of intelligent technologies in the field of library and information science include the following:

- cataloguing
- subject indexing
- collection management
- reference services including:
  - — referral of users to appropriate information resources
  - — selection of an appropriate database for searching information to meet a specific information need
- text processing including:
  - — text categorisation
  - — text summarisation
  - — intelligent agents for text processing
  - — text mining, data mining and knowledge discovery
- user interfaces.

Some of the applications of intelligent techniques mentioned above, such as in the area of collection management, must have in some form or another, intelligent techniques for processing and retrieval of information for specific tasks.

5.9 SUMMARY

- Search tactics is concerned with adopting techniques in manoeuvring the actual search.
- Guided search is designed for novice to intermediate searchers.
- Command search is designed for intermediate to experienced Dialog searchers.
- Data retrieval model essentially handles data.
- Knowledge is a kind of integration of general types of information.
- Based on theories and methods/tools available in other disciplines, a number of models have been developed in order to find satisfactory solutions for information retrieval problems.
- Fuzzy retrieval has its maximum utilisation in a system that accepts items that have been optical character recognised. In the OCR process, a hardcopy item is scanned into a binary image.
- The set theoretical view of information retrieval is based on the recognition that information requests are normally formulated by choosing collections or sets of item identifiers, or keywords.

- An intelligent information retrieval system is a system that carries out intelligent retrieval.

5.10 REVIEW QUESTIONS

NOTES

1. Give three examples of online search services.
2. What the major steps in an online information search?
3. What are the different Information Retrieval techniques?
4. Discuss briefly the various Retrieval Models.
5. What are the components of Expert Systems?
6. What is meant by intelligent information retrieval?
7. How does intelligent information retrieval differ from a conventional information?

5.11 FURTHER READINGS

1. Stacey, Alison, Stacey, Adrian. (2004). Effective Information Retrieval from the Internet. Oxford: Chandos Publishing.
2. Chowdhury, G.G. (2004). Introduction to Modern Information Retrieval. 2nd ed. London: Facet.
3. Fenichel, C.H. (1980). The Process of Searching Online Bibliographic Databases: A Review of Research. Library Research. 107–127.
4. Guinchat, P. and Menou, M. (1990). Sciences et techniques de I' information et de la documentation: introduction generale. Paris: Unesco.
5. Harter, Stephen P. (1988). Online Information Retrieval: Concepts, Principles and Techniques. London: Academic Press.