UNIT 9  BASICS OF SUBJECT INDEXING

Structure

9.0 Objectives

9.1 Introduction

9.2 Subject Indexing: Origin and Development

9.3 Meaning and Purpose

9.4 Cataloguing Versus Indexing

9.5 Indexing Principles and Process
  9.5.1 Need and Purpose of Indexing Principles
  9.5.2 Indexing Principles
  9.5.3 Indexing Policy
  9.5.4 Indexing Process
  9.5.5 Indexing Language
  9.5.6 Problems in Indexing
  9.5.7 Quality in Indexing

9.6 Evaluation of Indexing
  9.6.1 Purpose of Evaluation
  9.6.2 Efficiency and Effectiveness of Indexing System
  9.6.3 Evaluation Criteria
  9.6.4 Major Retrieval Experiments

9.7 Summary

9.8 Answers to Self Check Exercises

9.9 Keywords

9.10 References and Further Reading

9.0 OBJECTIVES

After reading this Unit, you will be able to:

- discuss the origin and development of subject indexing;
- explain the concept and purposes of subject indexing;
- discuss the principles of indexing;
- explain the different stages of intellectual operations in the subject indexing process;
- analyse and discuss the various problems associated with subject indexing in order to arrive at appropriate solutions;
- explain the reasons for evaluation of an indexing system; and
- discuss the major criteria for evaluation of the indexing systems.

9.1 INTRODUCTION

All library work is a matter of storage and retrieval of information, and cataloguing and indexing are specially performed to achieve that. Subject cataloguing is intended to
embraces only that activity which provides a verbal subject approach to materials added to library collections. Subject indexing is used in information retrieval especially to create index records to retrieve documents on a particular subject. Descriptive cataloguing makes it possible to retrieve the materials in a library by title, author, etc. — in short all the searchable elements of a cataloguing record except the subjects.

Until the second half of the nineteenth century, descriptive cataloguing was the basic library cataloguing practice that was found necessary. Libraries were much smaller than they are today, and scholarly librarians then were able, with the aid of printed bibliographies, to be familiar with everything available on a given subject and guide the users to it. With the rapid growth of knowledge in many fields during the nineteenth century and the resulting increase in the volume of books and periodicals, it became desirable to do a preliminary subject analysis of such works and then represent them in the catalogue or in printed indexes in such a way that they could be retrievable by subject. Subject cataloguing deals with what a book or other library item is about, and its purpose is to list, under one uniform word or phrase all the materials on a given topic that a library has in its collection. A subject heading is a uniform word or phrase used in the library catalogue to express a topic. The use of authorised words or phrases only, with cross-references from unauthorised synonyms, is the essence of bibliographic control in subject cataloguing. In the literature of LIS, the phrases subject cataloguing and subject indexing are used more or less interchangeably. In this context it is to be pointed out here that it was Charles Ammi Cutter who first gave a generalised set of rules for subject indexing in his Rules for a Dictionary Catalogue (RDC) published in 1876. But he never used the term ‘indexing’; rather he used the term ‘cataloguing’. In this course material, the phrase subject indexing includes subject cataloguing also. The literature differentiates the two as subject cataloguing is intended to embrace only that cataloguing activity which provides a verbal subject approach to library collections, especially macro documents (i.e. books). It refers to the determination and assignment of suitable entries for use in the subject component of a library’s catalogue. The primary purpose of the subject catalogue is to show which books on a specific subject are possessed by the library. Subject indexing refers to that indexing activity which provides a verbal subject approach to micro documents (e.g., journal articles, research reports, patent literature, etc.). Subject indexing provides a subject entry for every topic associated with the content of a micro document.

The representation of documents and the knowledge expressed by them is one of the central and unique areas of study within Library and Information Science (LIS) and is commonly referred to as indexing. Subject approach to information has been a long and extensive concern of librarianship and is assumed to be the major approach (access method) of users for a very long period. Indexing has traditionally been one of the most important research topics in information science. Indexes facilitate retrieval of information in both traditional manual systems and newer computerised systems. Without proper indexing and indexes, search and retrieval are virtually impossible.

### 9.2 SUBJECT INDEXING: ORIGIN AND DEVELOPMENT

The origin and development of subject indexing is intimately related with the historical development of libraries through ancient and medieval periods to modern days. The libraries of the ancient world used to arrange documents under some subjects. The catalogue, which worked as an index to this store, was predominantly a systematic subject listing according to a scheme of subject headings. The arrangement more or less conformed to the arrangement of documents in the store.
The specific usage of the term index goes back to ancient Rome. There, when used in relation to literary works, the term index was used for the little slip attached to papyrus scrolls on which the title of the work (and sometimes also the name of the author) was written so that each scroll on the shelves could be easily identified without having to pull them out for inspection. From this developed the usage of index for the title of books. In the first century A.D., the meaning of the word was extended from “title” to a table of contents or a list of chapters (sometimes with a brief abstract of their contents) and hence to a bibliographical list or catalogue. Only the invention of printing around 1450 made it possible to produce identical copies of books in large numbers, so that soon afterwards the first indexes began to be compiled, especially those to books of reference. By the end of 13th century alphabetisation by names of authors under systematic subject arrangement was well known. The index to the store or the shelf-list used to be supplemented with an author index to satisfy the author approach of the users of store. Index entries were not always alphabetised by considering every letter in a word from beginning to end. Most early indexes were arranged only by the first letter of the first word, the rest being left in no particular order at all. Gradually, alphabetisation advanced to an arrangement by the first syllable, that is, the first two or three letters, the rest of an entry still being left unordered.

The 15th century saw the entry of the university libraries which brought about a qualitative change. Efforts were made to rank subjects and on devising indexing or cataloguing methods for better utilisation of documents. Towards the end of 15th century the practice of supplementing systematic listing with an alphabetical subject index was introduced. Only very few indexes compiled in the 16th and early 17th centuries had fully alphabetised entries, but by the 18th century full alphabetisation became the rule. Alphabetical indexing gained a new momentum as intellectual debates among the scholars required ready reference to scholarly works with the rise of universities. The pressmarks, which were mainly used for storage of documents, started being used in catalogues as a retrieval tool. But the pressmarks could not ensure a flexible hierarchical order of subjects and hence it was discarded in favour of notation. In the 19th century, subject access to books was provided by means of classification. Books were arranged by subject and their surrogates were correspondingly arranged in a classified catalogue. Only in the late 19th century, alphabetical subject indexing became widespread and more systematic. Classification system was primitive in nature. It could not go deep enough to the extent of individualising subjects of documents. The separate existence of the classed catalogues and indexes stirred up imagination for compilation of a catalogue which was very much akin to a dictionary in form. Thus were born the forerunners of our dictionary and classified catalogues.

Preparation of back-of-the-book index, historically, may be regarded as the father of all indexing techniques. Indexing techniques actually originated from this index. 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 dictionary catalogue brought some relief into the sharp conflict between subjects of documents and the practice of naming them. Charles Ammie Cutter, who was both a classificationist as well as a theoretician of the library catalogue, observed that the name of the subject assigned to a document did not indicate its specific subject. Rather it indicated the class to which the subject of the document belonged to. For example assigning the subject ‘plant’ to a document discussing the plant ‘cactus’. The practice
Indexing

was deficient in helping a user who came for information on a specific subject. The root of the conflict remained deep in the classification system also as the classification was not coextensive with the subjects of the documents. Hence, whatever was left out in classification became conspicuous by their absence while giving class names to individual entries as subject heading. Cutter, who was an advocate of dictionary catalogue wanted to solve the conflict at the cataloguing level. The year 1876 is particularly important for the library profession for the publication of two outstanding books: (1) *A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library*, by Melvil Dewey; and (2) *Rules for a Dictionary Catalogue*, by Charles Ammie Cutter. The first sought to solve the problems by organising the document store and simultaneously providing an alphabetical subject index for easy access to it while the second, expressing doubts about the efficacy of class headings to be used as specific subject heading, decided to go through a different way by some specific method for naming of the subjects. While Dewey offers a ready made list of names (class names in this case) Cutter suggested some methods for building them up in order to name them more specifically. Cutter’s rules for specific subject headings for use in a dictionary catalogue seemed to have appealed the library professional. Subsequently there was a demand for some ‘standard list of subject headings’ which could be used in carrying out the specifications in Cutter’s rules. This paved the way for the publication of a list of subject headings by the American Library Association (ALA), to be used in a dictionary catalogue. The list was later revised and published in two more editions which ultimately established a pattern for subsequent subject heading lists like subject headings used in the dictionary catalogue of Library of Congress and Sears List of Subject Headings.

Use of the above standard lists of subject headings raised important questions relating to the use of terminology (whether common or popular terms or scientific and technical terms were to be used), and sequencing of terms in the subject heading (what should be the sequence of terms in case of compound subject headings). But Cutter as well as compilers of several standard lists of subject headings failed to provide satisfactory answers to the above noted questions.

The first quest for logical approach towards solving the above noted problems is evident in J. Kaiser’s *Systematic Indexing* (1911). Kaiser was the first person who gave the idea of categorising the terms under two fundamental categories: concretes and processes. He recommended the citation order of these categories into index string. Kaiser suggested that many composite subjects could be analysed into a combination of concepts indicating a ‘concrete’ object and a ‘process’. In such cases the concrete should be given precedence over ‘process’ in the order of citation of index terms in a compound subject heading. Kaiser failed to analyse deeply the various types of intricacies involved in naming of subjects. Nevertheless, his work remains unique till date as he is the first person to suggest certain logical processes for naming subjects in terms of fundamental categories and a citation order of index strings.

Dr. S. R. Ranganathan was the first to analyse the universe of subjects in depth and suggesting a complete theory of naming subjects using a subject indexing language. He realised the fallacy of trying to symbolise the extremely flexible and dynamic multidimensional universe of subjects into a linear, rigid notational model. Just as ready-made class numbers cannot be given according to his scheme of classification to all subjects of the past, present and future, so also subject headings cannot be made available ready-made. He therefore enunciated certain rules on the basis of which subject names could be framed. Ranganathan developed a mechanical procedure for doing it and called it the chain procedure. The basic contention of chain procedure is that a multidimensional universe of subjects cannot be fitted into a rigid one-dimensional model
and hence, a chain of terms is required to name a subject where the term indicating the specific subject is stated in a particular context. Chain procedure demonstrated that it is not necessary to depend on the flair of some authorities for supply of names of subjects. One can very well build up one’s own authority file and use subject names consistently. The names used will be uniform for all libraries following the same scheme of classification. The chain, which is a string of terms, gets organised or arranged following the classification scheme used. Qualities of the classification scheme therefore very much determine the qualities of the subject headings drawn according to chain procedure.

J. E. L. Farradane devised a scheme of pre-coordinate indexing system known as Relational Indexing in the early 1950s. The basic proposition of Farradane’s Relational Indexing was to identify the relationship between concepts by following the learning process through which we develop our power of discrimination in time and space. Farradane’s Relational Indexing has been the subject of scholarly research, but was never implemented. Still we can say that Farradane’s contribution to the area of subject indexing was: analysis of relationship among each pair of terms, use of relational operators, and representation of relationship among terms by relational operators leading to the creation of ‘Analets’. ‘Analet’ refers to a pair of terms linked by any of the relational operators as developed by Farradane. Each relational operator is denoted by a slash and a special symbol having a unique memory. For example,

Chemistry / \( \Theta \) Encyclopaedia

The contribution of E. J. Coates in subject indexing was not original in nature. Coates merely synthesized the ideas of Cutter, Kaiser, Ranganathan and Farradane. Coates applied his idea on British Technology Index (now Current Technology Index) of which he was editor from its inception in 1963 until his retirement in 1976.

Preserved Context Index System (PRECIS), developed by Derek Austin and applied to BNB in 1971 as an alternative to the chain procedure for deriving subject index entries, sought to rectify the problem of co-extensiveness by generating entries with a lead term and the full context of the document. Depending heavily on computer to generate mechanically all index entries from input strings, PRECIS developed its own code for preparation of input strings by the human indexer and its subsequent processing by computer. Its emphasis has been on generating printed index for BNB. Though PRECIS was fairly successful in its original mission it does not have the simplicity of chain procedure and considerable skill is required to use it effectively.

Postulate-based Permutated Subject Indexing (POPSI) sought to overcome the shortcomings of chain procedure from entirely different perspective. It recommended postulates and principles for analysing the subjects into elementary categories and their subsequent ordering. The postulates are not rigid and hence give flexibility to indexers. As it is essentially distilled out of chain procedure it has managed to retain most of the helpful features of chain procedure such as simplicity. Over the years, Bhattacharya,
Neelameghan, Devadasan, Gopinath and others have given a sound theoretical foundation to POPSI in terms of ‘General Theory of Subject Indexing Languages’ (GT-SIL). The GT-SIL seeks to analyse the deep structure of Subject Indexing Languages in terms of semantic structure, elementary structure and syntactic structure of subject propositions. In essence, GT-SIL is a logical abstraction of the structures of outstanding subject indexing languages such as those of Cutter, Dewey, Kaiser and Ranganathan.

It is evident from the above discussion that the research on development and use of various subject indexing systems was devoted to techniques of constructing pre-coordinate subject headings. A greater part of the pre-coordinate subject indexing system was devoted to syntactical rules of indexing. Rigidity of significance order may not meet the approaches of all users of the index file, though this problem is solved by rotating terms or multiple entry system. It is also evident that even the acceptance of multiple entry system covers only a fraction of the possible number of total permutations. Thus, a large portion of probable approaches or access points are left uncovered. This gap widens rapidly with every increase in the number of terms in a subject heading due to the demand for more specific subject headings. The index file may fail to provide a particular combination which the user is looking for. It may also provide a combination which proves too broad for a particular search. The above considerations and difficulties stemming from the pre-coordination of terms led to the development of post-coordinate indexing or simply coordinate indexing systems like Uniterm, Optical Coincidence Card / Peek-a-boo, Edge-Notched Card, etc during 1960s.

Computers began to be used to aid information retrieval in the 1950s. The Central Intelligence Agency (CIA) of USA is said to be the first organisation to use the machine-produced keywords from Title Index 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 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’. A number of varieties of keyword index are evident in the literature. They differ only in terms of their formats but indexing techniques remain more or less the same.

The publication of Science Citation Index (SCI) by Eugene Garfield of the Institute of Scientific Information (ISI), Philadelphia in 1963 provided a new approach to the bibliographic file organisation. The online version of the SCI, known as SCISEARCH, was published in 1974. ISI also brought out the Social Science Citation Index (SSCI) and Arts and Humanities Citation Index (A&HCI) in 1973 and 1978 respectively. The publication of the citation classics, with the first issue of Current Contents in 1977, forms an important and interesting venture of the ISI.

It has already been mentioned above that the traditional subject indexing systems and techniques have taken a new turn with the applications of computers in 1950s. In fact, all attempts at computerised indexing were based on two basic methods: Statistical analysis; and Syntactic and semantic analysis. In the arena of computerised indexing, there has been considerable research on the user-interface design, indexing systems using Artificial Intelligence techniques like Natural Language Processing (NLP), Knowledge Representation Model and Expert System-based subject indexing systems. As a result of the phenomenal growth of content on the web as an indexing problem, we have seen a continued interest in the development of tools and techniques to index the Web resources. Different search tools and technologies were developed in finding
9.3 MEANING AND PURPOSE

The term index came from the Latin word indicare which means ‘to point out, to guide, to direct, to locate’. An index indicates or refers to the location of an object or idea. It is a methodically arranged list of items or concepts along with their addresses. The process of preparing an index is known as indexing. According to the British Standards (BS 3700: 1964), an index is “a systematic guide to the text of any reading matter or to the contents of other collected documentary material, comprising a series of entries, with headings arranged in alphabetical or other chosen order and with references to show where each item indexed is located”. An index is, thus, a working tool designed to help the user to find his way out of a 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, audio-visual documents, and computer-readable records including web resources.

It appears from the foregoing discussion that an index indicates or refers to the location of an object/idea/concept. A concept is a unit of thought. The semantic content of a concept can be re-expressed by a combination of other and different concepts, which may vary from one language or culture to another. What the particular body of information is about, in a document constitute its subject. A subject can be defined as any concept or combination of concepts representing a theme in a document. An indexing term is defined as the representation of a concept in the form of either a term derived from natural language or a classification symbol.

A subject is then any concept or combination of concepts which is expressed in the document. The readers’ task is to interpret the words and sentences in the document in order to understand the concepts. Whether a reader understands a document depends on how precisely the author expresses the concepts he refers to and whether the reader is aware of the concepts the author expresses. The basic idea is that the concepts exist before the author writes the document and the reader reads the document.

Similarly, the indexer’s task is to identify concepts in the document and re-express these in indexing terms. This is done first by establishing the subject content, or in other words the content of concepts in the document. Thereafter the principal concept presented in the subject content is identified, and finally, the concepts are expressed in the indexing language. The indexing is successful when the document and the indexing term express the same concepts.

• Purpose

Modern subject indexing practice has its roots in Charles Ammi Cutter’s Rules for a Dictionary Catalog published in 1876. In chapter 1 Cutter’s statement of the basic objectives of a catalogue is: (i) To enable a person to find a book of which the subject is known, and (ii) To show what the library has on a given subject (and related subjects). This implies that the main purpose of subject indexing is to satisfy the subject query of the users by enabling an enquirer to identify documents on a given subject and providing information on the presence of material on allied or related subjects.

The first objective refers to the need to locate individual items, and the second refers to the need to collocate materials on the same subject as well as related subjects. A subject is a set of interrelated component ideas in which each component idea is related
Indexing directly or indirectly to other component ideas. A subject of a document is amenable for structuring into subject heading. It is a kind of linear structuring of subject surrogates, and some criteria for formatting or modeling it into an accessible procedure. The purpose of subject indexing is to:

a) satisfy the subject approach to information;
b) identify pertinent materials on a given subject or topic;
c) enable the enquirer to find materials on related subjects;
d) link related subjects by a network of references;
e) prescribe a standard methodology to subject cataloguers/indexers for constructing uniform subject headings;
f) bring consistency in the choice and rendering of subject entries, using standard vocabulary and according to the given rules and procedures;
g) be helpful to users in accessing any desired document(s) from the catalogue or index through different means of such approach;
h) decide on the optimum number of subject entries, and thus economize the bulk and cost of indexing; and
i) provide user-oriented approach in naming the subjects through any vocabulary common to a considerable group of users, specialists or laymen.

Self Check Exercise

Note:  

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

1) What do you mean by ‘Subject Index’ and ‘Subject Indexing’?

2) State the purposes of subject indexing.

3) Traces the historical development of subject indexing.
There has always been confusion about the distinction between ‘subject cataloguing’ and ‘subject indexing’. Basically, cataloguing is the process of creating bibliographic description of a document as a whole entity, and subject cataloguing and classification assign subject labels which together describe the overall topic of the document. Indexing involves delving into a document for analysing its contents at a much deeper level to provide access to many of the concepts contained within it at greater depth. Although most articles in a periodical issue and many books are listed under only one or two subject headings, a back-of-the-book index contains hundreds of subject terms associated with the content of an individual book.

Subject cataloguing usually refers to the assignment of subject headings to represent the overall contents of whole documents (e.g. books, reports, periodicals, etc.) within the catalogue of a library. Subject indexing is a term used more loosely; it may refer to the representation of the subject matter of the parts of whole documents as in the case of a back-of-the-book subject index. Thus, a library may enter a book under the subject heading ‘roses’ in its catalogue to indicate its overall subject matter, the detailed contents of the book are only revealed by the back-of-the-book subject index. This distinction between the terms ‘subject cataloguing’ and ‘subject indexing’, one referring to complete bibliographic items and other to parts of them, is artificial, misleading, and inconsistent. The process by which the subject matter of documents is represented in databases—printed or electronic form—is almost referred to as ‘subject indexing’, whether overall documents or their parts are discussed. Thus, the subject index might refer to the representation of the content of the complete books or complete technical reports as well as to the parts of documents (e.g. chapters in books, papers within the periodicals or conference proceedings, etc.). On the other hand, libraries may choose to represent parts of books (e.g. chapters or papers) within the catalogue which is usually referred to as analytical cataloguing.

The situation is even more confusing when the term classification is considered. The term classification refers to the process of assigning class numbers, drawn from a given classification scheme, to documents, especially for the purpose of arranging these items on the shelves of the libraries, in catalogues, etc. But the subject catalogue of a library can be either alphabetically based (in an alphabetical subject catalogue or dictionary catalogue) or arranged according to the sequence of a classification scheme (in a classified catalogue). Suppose a librarian picks up a book and decides that it is about ‘banking’. He or she might assign the subject heading Banking to this document. Alternatively, the Dewey Decimal classification number 332.1 may be assigned to it. Many people would refer to the first operation as subject cataloguing and to the second as classification. These terminological distinctions are quite meaningless and only serve to create confusion due to failure to understand the distinction between the conceptual analysis and translation stage in indexing. In short, subject indexing is conceptually identical to subject cataloguing. Its process involves classification, forming classes of objects on the basis of their subject matter and representing them either in the verbal plane (by using a readymade list of subject headings or a thesaurus) or in the notational plane (by using a scheme of classification). In this Unit, the term subject indexing or simply indexing is used as a matter of convenience to refer to all activities of subject cataloguing.
9.5 INDEXING PRINCIPLES AND PROCESS

9.5.1 Need and Purpose of Indexing Principles

Before we discuss the principles of indexing, it is important to know why we need to have principles of indexing. We need principles of indexing:

1) To set out the general directions for the consistent application of subject indexing techniques;
2) To serve as a useful guide for developing new indexing techniques and to develop one that already exists;
3) To facilitate the evaluation of indexing systems;
4) To provide theoretical rationale for particular standards or guidelines for designing subject indexing system and its application;
5) To promote understanding of different subject indexing systems by identifying commonalities underlying them and providing a structure for their comparison; and
6) To determine how the subject headings are established and applied.

9.5.2 Indexing Principles

Indexing principles may be stated as:

a) **The user as focus:** The wording and structure of the subject heading should match what the user will seek in the index;

b) **Unity:** A subject index must bring together, under one heading all the documents which deal principally or exclusively with the subject, whatever the terms, applied to it by the authors and whatever the varying terms, applied to it at different times. It must use a term which is unambiguous and does not overlap in meaning with other headings in the index.

c) **Common Usage:** The subject heading chosen must represent common usage or, at any rate, the usage of the class of users for whom the documents on the subject within which the heading falls is intended. Whether a popular term or a scientific one is to be chosen should depend on the approaches of the users.

d) **Specificity:** The heading should be as specific as the topic it is intended to cover. As a corollary, the heading should not be broader than the topic. Rather than using a broader heading, the cataloguer should use two specific headings which will approximately cover it.

9.5.3 Indexing Policy

Indexers must take policy decisions about how many terms should be included in an index entry, how specific the terms should be and how many entries an index should incorporate. Together this gives a depth of indexing. The depth of indexing describes the thoroughness of the indexing process with reference to exhaustivity and specificity. While taking such a policy decision, indexers should strive for a balance between specificity and exhaustivity and should consider the requirement of the users of the index along with the cost and time factors.
Exhaustivity in Indexing

Exhaustivity in indexing is the detail with which the topics or features of a document are analysed and described. In other words, an exhaustive index is one which lists all possible index terms associated with the thought content of a document. In contrast to higher exhaustivity, higher specificity increases precision at the cost of impaired recall. Greater exhaustivity gives a higher recall leading to the retrieval of all the relevant documents along with the retrieval of a large number of irrelevant documents or documents which only deal with the subject in little depth.

Specificity in Indexing

The specificity describes how closely the index terms match the topics they represent in a document. It is the extent to which the indexing system permits us to be precise when specifying the subject of a document we are processing. Higher specificity leads to high precision, whereas lower specificity will lead to low precision, but high recall. Specific indexing provides specific terms for all or most topics and features and results in a larger indexing vocabulary than more generic indexing. Specificity tends to increase with exhaustivity in indexing vocabulary as the more terms we include, the narrower those terms will be. A high level of specificity increases precision.

9.5.4 Indexing Process

The representation of documents and the knowledge expressed by them is one of the central and unique areas of study within library and information science (LIS) and is commonly referred to as indexing. A common demand in the LIS field is for a set of rules or a prescription for how to index. When this demand is raised it is usually based on the assumption that it is possible to explain the intellectual operations in the subject indexing process. The indexing process basically consists of two intellectual steps: conceptual analysis and translation.

Conceptual analysis

This step refers to the identification of different component ideas associated with the thought content of the document and establishment of interrelationship between those component ideas. According to Ranganathan, it involves the work in the idea plane which is carried out in two stages, although these tend to overlap in practice:

a) examining the document and establishing its subject content;

b) identifying the principal concepts present in the subject;

- Examinining the document and establishing its subject content

In the first stage of the conceptual analysis of the thought content of the document, it is examined for the establishment of its subject content. A complete reading of the document often is impracticable, but the indexer should ensure that no useful information has been overlooked. While examining the document, the indexer should give particular attention to a number of places in the document: the title; the abstract, if provided; the list of contents; the introduction, the opening chapters and paragraphs, and the conclusion; illustrations, diagrams, tables and their captions; words or groups of words which are underlined or printed in an unusual typeface.

- Identifying the principal concepts present in the subject

In this stage of the indexing process the indexer identifies the principal concepts in the subject. The second stage is laid over the first stage in the sense that the indexer should
not go back to the document to look for concepts. Rather, the indexer should look for concepts within the findings of the first step; that is the natural language representations of the subject content. The indexer does not necessarily need to retain, as indexing elements, all the concepts identified during the examination of the document. After examining the document, the indexer needs to follow a logical approach in selecting those concepts that best express its subject. While selecting the principal concepts of the document the indexer should take into consideration the purpose for which the indexing data will be used. Indexing data may be used for the purpose like preparation of subject headings for the subject catalogue, production of printed alphabetical indexes to different types of information products, and computerised storage of indexing data elements for subsequent retrieval of the documents.

**Translation**

During the first two stages the indexer has established the subject content of the document and identified the principal concepts in the subject. The indexer is hereafter ready to translate the concepts into the indexing language. This step refers to the expression of principal concepts as identified while analysing the thought content of the document into the language of the indexing system. According to Ranganathan, it involves the work in the verbal plane which calls for the familiarity with different components of the given indexing language: controlled vocabulary, syntax and semantics including their working roles for displaying the indexing data in a subject index.

If the concepts that the indexer has identified during the second stage are present in the indexing language the indexer should translate the concept into preferred terms. At this point in the indexing process the indexer should be aware that indexing languages may impose certain constraints in translating the concepts. If the indexer uses a controlled indexing language, this may not permit the exact representation of a concept encountered in a document. The concern is that the concepts that the indexer identified during the second stage of the indexing process might not be present in the indexing language. The indexer is then forced either to choose a term that does not express exactly the same concept or add a new term to the vocabulary to represent the concept. Here, the indexer is required to be familiar with the particular indexing language and the specific rules and mechanisms of the indexing language.

### 9.5.5 Indexing Language

An indexing language is a set of terms and devices used to establish the relationship between terms for representing the content of the documents as well as queries of the users. It consists of three basic elements: controlled vocabulary, syntax and semantics. Controlled vocabulary has been defined as a limited set of terms showing their relationships and indicating ways in which they may usefully be combined to provide subject index to the documents and to search for these documents, in a particular system. Syntax comprises a grammatical structure or a set of rules that govern the sequence of occurrence of terms/words in representing the content of the document. Semantics refers to the systematic study of how meaning is structured, expressed and understood in the use of an indexing language. More discussion on indexing languages can be seen in the Unit 10 of this Course.

### 9.5.6 Problems in Indexing

An indexer analyses a text and strives to ascertain meaning. Ideally this analysis anticipates a searcher at some future time, looking for text with the same meaning. But, meaning is not fixed at either end of this process. And even if the meaning is relatively unambiguous
or stable, the terms used to represent it are not. Thus, most indexing processes encounter problems at two levels:

- Interpreting meaning as intended by the author and as construed by the potential user;
- Choosing the terms to represent that meaning that will enable this communication to be clear and as true as it can be. (Bearing in mind that such fidelity is a relative thing to begin with)

Fidelity in the context of IR denotes the accuracy with which term(s) used to represent the name of the subject represent the meaning. A number of problems and issues associated with indexing are:

a) Subjects of documents are complex—usually multi-worded terms;

b) Users’ request for information tend to multidimensional;

c) Choice of terms—among different categories, viz. entities, activities, abstracts, properties and heterogeneous concepts (synonymous to semantic factoring);

d) Choice of word forms—among different forms, viz. noun vs. adjective, singular vs. plural;

e) Homographs—if neglected, will give rise to reduced relevance. Seriousness of the problem will depend on the coverage of the system.

f) Choice of the kind of vocabulary that should be used, and syntactical and other rules necessary for representing complex subjects;

g) Identification of term relationship—semantic vs. syntactic;

h) Decision about the exhaustivity level (i.e. the depth to which indexing should be done);

i) Decision about the specificity level (i.e. The levels of generality and specificity at which concepts should be represented);

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

k) 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 expected to benefit from it and the types of requests for which the document is likely to be regarded as useful.

### 9.5.7 Quality in Indexing

The quality of an index is defined in terms of its retrieval effectiveness—the ability to retrieve what is wanted and to avoid what is not. Quality in indexing leads to a better performance in retrieving documents. The governing idea is that indexing should be neutral, objective, and independent of the particular indexer’s subjective judgment. An indexing failure on the part of the indexer may take place at the following stages of indexing process:

- Failure in establishing concepts during conceptual analysis of the content of a document;
Indexing

- Failure to identify a topic that is of potential interest to the target user group;
- Misinterpretation of the content of the document, leading to the selection of inappropriate term(s);
- Failure in translating the result of conceptual analysis into the indexing language;
- Failure to use the most specific term(s) to represent the subject of the document;
- 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
- Omission of important term(s).

The quality of indexing depends on two factors: (i) qualification and expertise of the indexer; and (ii) quality of the indexing tools. In order to achieve quality in indexing, the indexer should have adequate knowledge of the field covered by the documents s/he is indexing. S/he should understand the term of the documents as well as the rules and procedures of the specific indexing system. Quality control would be achieved more effectively if the indexers have contact with users. An indexer who has contact with the users might better be able to represent the documents in accordance with how the users think. The idea is that the indexer should attempt to determine the subject of the document taking into account the users’ questions and information needs. This might help the indexer when a document contains multiple concepts. In such a situation, the indexer can select only those concepts to represent the content of a document which are regarded as most relevant by a given community of users. Indexing quality can be tested by analysis of retrieval results, e.g. by calculating recall and precision ratios.

Indexing Consistency

It is assumed that there is a relationship between indexing consistency and the indexing quality. That is to say, an increase in consistency can be expected to cause an improvement in indexing quality. Traditionally, consistency in indexing has long been considered as an acceptable indicator of indexing quality. Consistency in indexing is essential for effective retrieval. Indexing consistency refers to “the extent to which agreement exists on the terms to be used to index some document” (Lancaster, 2003). 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 in the indexes. The goal of the consistency is to promote standard practice in indexing.

It has for long been observed that different indexers tend to assign different index terms to the same document as they differ considerably in their judgment as to which terms reflect the contents of the document most adequately. Essentially, indexing consistency is seen as a measure of the similarity of reaction of different human beings processing the same information. Indexing consistency in a group of indexers is defined as the degree of agreement in the representation of the essential information content of the document by certain sets of indexing terms selected individually and independently by each of the indexers in the group.

In the process of indexing, indexers choose what topics to represent and what to call those topics. The goal is to select and name topics consistently so that all of the material about any given topic will be found together. Ideally, if two indexers use the same thesaurus or classification system to index the same document, they are supposed to assign the same index terms or class numbers. In practice, indexers are not always
consistent with each other, because subject indexing is essentially a subjective process. Indexers may miss important points of the document, and add irrelevant terms. This would stem from insufficient knowledge of indexers about the subject. Decades of research on consistency between indexers and by the same indexer at different times has documented medium to high levels of inconsistency.

Self Check Exercise

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

4) Explain the steps involved in subject indexing.

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

5) What are the different levels of indexing principles?

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

6) Distinguish between Exhaustivity and Specificity in Indexing.

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

7) Identify the different stages in which indexing failure can take place.

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

9.6 EVALUATION OF INDEXING

An indexing system is a sub-system of an information retrieval system and hence, its performance is directly linked up with the overall performance of the entire information retrieval system.

Evaluation of an indexing system essentially means measuring the performance of the system, success or failure, in terms of its retrieval efficiency (ease of approach, speed and accuracy) to the users, and its internal operating efficiency, cost effectiveness and cost benefit to the managers of the system.
The foundation of the Institute of Information Scientists in the UK in 1958 coincides closely with the beginning of the notion of experimental evaluation of information retrieval systems in general and indexing system in particulars. Although there had been some earlier attempts, we usually mark the start of the tradition as the Cranfield experiments, which ran from 1958 to 1966.

9.6.1 Purpose of Evaluation

- To identify the level of performance of the given indexing system,
- To understand how well the given indexing system fulfills the queries of the users in retrieving the relevant documents,
- To compare the performance of two or more indexing systems against a standard,
- To identify the possible sources of failure of the given indexing system or inefficiency with a view to raising the level of performance at some future date,
- To justify the existence of the given indexing system by analysing its costs and benefits,
- To establish a foundation of further research on the reasons for the relative success of alternative techniques, and
- To improve the means employed for attaining objectives or to redefine goals in view of research findings.

9.6.2 Efficiency and Effectiveness of Indexing System

By effectiveness we mean the level up to which the given indexing system attains its stated objectives. The effectiveness may be a measure of how far an information retrieval system can retrieve relevant information withholding non-relevant information. The effectiveness of an indexing system can be measured by calculation of recall and precision ratios. By efficiency we mean how economically the indexing system is achieving its stated objectives. Efficiency can be measured by such factors, such as, at what minimum cost and effort does the system function effectively. It may be necessary that the cost factors are to be calculated indirectly, such as response time (i.e. that is time taken by the system to retrieve the information), user effort (i.e. the amount of time and effort required by a user to interact with the indexing system and analyse the output retrieved in order to get the required information), the cost involved, and so on.

9.6.3 Evaluation Criteria

It is evident from the history of the experimental evaluation of information retrieval systems that there has been a remarkably coherent development of a set of criteria for the evaluation of indexing systems. These evaluation criteria generate argument, disagreement and heated dispute, but there remains a relatively stable common core, which has, despite its limitations, served us well over the last 50 years. The most important criteria used for evaluating an indexing system are: Recall and Precision.

Recall and Precision

- Recall

Recall refers to the index’s ability to let relevant documents through the filter. Recall ratio is a ratio of the relevant documents retrieved to the total number of relevant documents potentially available. It measures the completeness of the output. Hence,
the recall performance can be expressed quantitatively by means of a ratio called recall ratio as mentioned below:

\[
\text{Recall ratio} = \frac{R}{C} \times 100
\]

Where,  
R = Number of relevant documents retrieved against a search  
C = Total number of relevant documents available to that particular request in the collection.

- **Precision**

When the system retrieves items that are relevant to a given query it also retrieves some documents that are not relevant. These non-relevant items affect the success of the system because they must be discarded by the user, which results in wastage of a significant amount of time. The term ‘precision’ refers to the index’s ability to hold back documents not relevant to the user. Precision ratio is a ratio of the relevant documents retrieved to the number of documents retrieved. It measures the preciseness of the output, i.e. how precisely an indexing system functions. If recall is the measure of system’s ability to let through wanted items, precision is the measure of the system’s ability to hold back unwanted items. The formula for calculation of precision ratio is:

\[
\text{Precision Ratio} = \frac{R}{L} \times 100
\]

Where,  
R = Total number of relevant documents retrieved against a search  
L = Total number of documents retrieved in that search

The search result against a query is to separate the all documents into two parts: (a) One part is the set of relevant documents, and (b) the other part is the set of irrelevant documents. The following matrix can be used as a common frame of reference for evaluation of indexing system with reference to the calculation of recall and precision ratios:

<table>
<thead>
<tr>
<th>User relevance decision</th>
<th>Retrieved</th>
<th>Not retrieved</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>a</td>
<td>c</td>
<td>a + c</td>
</tr>
<tr>
<td></td>
<td>Hits</td>
<td>Misses</td>
<td>Total relevant</td>
</tr>
<tr>
<td>Noise</td>
<td>b</td>
<td>d</td>
<td>a + b</td>
</tr>
<tr>
<td>Total retrieved</td>
<td>a + b</td>
<td>c + d</td>
<td>a + b + c + d</td>
</tr>
<tr>
<td>Total not retrieved</td>
<td>a + c</td>
<td>b + d</td>
<td>Total collection</td>
</tr>
</tbody>
</table>

From the above matrix, recall and precision ratios can be calculated according to the following manner:

- Recall ratio = \([a / (a+c)] \times 100\)

- Precision ratio = \([a / (a+b)] \times 100\)

Where,  
\[a = \text{Hit (Retrieval of relevant documents by the system. It adds to precision).}\]  
\[b = \text{Noise (Retrieval of irrelevant documents by the system along with the relevant documents against a search).}\]
Indexing

c = Misses (The system fails to retrieve the relevant documents that should have been retrieved. It adds to the noise).

d = Dodged (The system correctly rejects to retrieve the documents that are not relevant to the given query).

It needs to be pointed out here that 100% recall and 100% precision are not possible in practice because recall and precision tend to vary inversely in searching. When we broaden a search to achieve better recall, precision tends to go down. Conversely, when we restrict the scope of a search by searching more stringently in order to improve the precision, recall tends to deteriorate.

Relevance

In human history, relevance has been around forever, or as long as humans tried to communicate and use information effectively. The concept of “relevance” is the fundamental concept of information science in general and information retrieval, in particular. Evaluation of indexing will never be effective until there is an understanding of the percept of relevance. Relevance is one of the important types of measures used in the evaluation of an information retrieval system and is highly debated issue in information retrieval research. There does not seem to be any consensus among the experts on the definition of relevance.

The first full recognition of relevance as an underlying notion came in 1955 with a proposal to use “recall” and “relevance” (later, because of confusion, renamed precision, sometimes it was called as pertinence) as measures of retrieval effectiveness in which relevance was the underlying criterion for these measures. But, the term pertinence refers to a relationship between a document and an information need, whereas the term relevance refers to a relationship between a document and a request statement (i.e. expressed information need). It refers to the ability of an information retrieval system to retrieve material that satisfies the needs of the user.

We know that the main objective of indexing, forming an essential component of an IR system, is to determine the aboutness of documents for subsequent retrieval of information object relevant to user queries. Relevance denotes how well a retrieved set of documents meets the information need of the user i.e. to what extent the topic of a retrieved set of information objects matches the topic of the query or information need.

In most of the evaluation studies relevance was applied to stated requests (i.e. expressed need). But, it has now been well established that the users’ requests do not reflect their information needs completely. Therefore, the current view is that the relevance is to be judged in relation to both expressed and unexpressed needs rather than restricting only to stated requests. It is dependent on the degree to which a user is able to recognize the exact nature of his/her information need and the degree to which his/her need is accurately expressed in the form of a request (i.e. request statement). Information retrieval systems create relevance—they take a query, match it to information objects in the system by following some algorithms, and provide what they consider relevant. People derive relevance from obtained information or information objects. They relate and interpret the information or information objects to the problem at hand, their cognitive state, and other factors—in other words, people take the retrieved results and derive what may be relevant to them. Relevance is derived by inference.

Although “relevance” is extensively used in evaluation of information retrieval, there are considerable problems associated with reaching an agreement on its definition, meaning, evaluation, and application in information retrieval. There are a number of different
views on “relevance” and its use for evaluation. This is because there are degrees of relevance. Relevance is a subjective factor depending on the individual. The same questions, posed by two different enquirers, may well require two different answers. It is because of the fact that enquirers seek information from their own corpus of knowledge. Thus it appears that the relevance is highly subjective and personal. It is a relation between an individual with an information need and a document.

**Other Important Criteria**

Perry and Kent are credited for bringing the concept of evaluation into information retrieval systems during 1950s. The evaluation criteria they suggested were:

i) **Resolution factor:** The proportion of total items retrieved over a total number of items in the collection.

ii) **Pertinency factor:** The proportion of relevant items retrieved over a total number of retrieved items. This factor was popularly named as the precision ratio in the subsequent evaluation studies.

iii) **Recall factor:** The proportion of relevant items retrieved over a total number of relevant items in the collection.

iv) **Elimination factor:** The proportion of non-retrieved items (both relevant and non-relevant) over the total items in the collection.

v) **Noise factor:** The proportion of retrieved items those are not relevant. This factor is considered as the complement of the pertinency factor.

vi) **Omission factor:** The proportion of non-relevant items retrieved over the total number of non-retrieved items in the collection.

Perry and Kent suggested the following formulae for the estimation of the above mentioned evaluation criteria:

\[
\frac{L}{N} = \text{Resolution factor} \quad \frac{(N-L)}{N} = \text{Elimination factor}
\]

\[
\frac{R}{L} = \text{Pertinency factor} \quad \frac{(L-R)}{L} = \text{Noise factor}
\]

\[
\frac{R}{C} = \text{Recall factor} \quad \frac{(C-R)}{C} = \text{Omission factor}
\]

Where,  

- \(N\) = Total number of documents
- \(L\) = Number of retrieved documents
- \(C\) = Number of relevant documents
- \(R\) = Number of documents that are both retrieved and relevant

C. W. Cleverdon (1966) identified six criteria for the evaluation of an information retrieval system. These are:

i) **Recall:** It refers to the ability of the system to present all the relevant items;

ii) **Precision:** It refers to the ability of the system to present only those items that are relevant;

iii) **Time lag:** It refers to the time elapsing between the submission of a request by the user and his receipt of the search results.

iv) **User Effort:** It refers to the intellectual as well as physical effort required from the
user in obtaining answers to the search requests. The effort is measured by the amount of time user spends in conducting the search or negotiating his enquiry with the system. Response time may be good, but user effort may be poor.

v) **From of presentation** of the search output, which affects the user’s ability to make use of the retrieved items, and

vi) **Coverage of the collection**: It refers to the extent to which the system includes relevant matter. It is a measure of the completeness of the collection.

9.6.4 Major Retrieval Experiments

- **Cranfield Test**

The first extensive evaluation of retrieval systems was carried out by the ASLIB with financial assistance from the National Science Foundation at the College of Aeronautics, Cranfield, UK, under the supervision of C. W. Cleverdon in 1957. It is popularly known as the Cranfield 1 Project. Its objective was to investigate into the comparative efficiency of four indexing systems: UDC, Faceted Classification, Alphabetical Subject Heading List and Uniterm Indexing system. Cranfield 1 was the most significant study in the decade of 1958-68. Even then it was not free from criticisms. One of the main criticisms against this study was the artificiality without much relation to the real life situation. The drawbacks of Cranfield 1 necessitated the conduct of further tests. The second stage of Cranfield studies, known as Cranfield 2, began in 1963 and completed in 1966.

- **MEDLARS Test**

The largest evaluation of an operating system was performed by Lancaster on the performance of the Medical Literature Analysis and Retrieval System (MEDLARS) of the US National Library of Medicine in the period between August 1966 and July 1967. The study involved the derivation of performance figures and conduct of details of failure analyses for a sample of 3000 real searches conducted in 1966-67. The objective of the MEDLARS test was to evaluate the existing system and to find out how it could be improved. The document collection available on the MEDLARS service at the time of the test consisted of about 700,000 items. Search requests from twenty-one user groups, selected for the study were taken. On receipt of a request, the system operator formulated it in **Mesh** (Medical Subject Headings) terms using expansion and search logic.

- **TREC Experiment**

Prior to 1991, research in information retrieval was limited to small test collections, and there was little transfer of research ideas into commercial systems. The Text RETrieval Conference (TREC), launched in 1991, introduced the first large test collection of full-text documents in order to enable IR researchers to scale up from small collection of data to larger experiments, along with the idea that providing an open testing event, with common tasks and a standard evaluation scenario, would lead to the acceleration of research on a realistic scale. It was funded by DARPA (Defence Advanced Research Project Agency, Department of Defence, USA) and operated by the NIST (National Institute for Science and Technology, USA). It is an annual competition/collaboration/get-together between research groups interested in different aspects of information retrieval. Every year, a set of tasks is defined, broadly information retrieval/search tasks. The TREC series of experiments has drawn attention of the LIS professionals all over the world since its inception and has shown that significant research results can be obtained through international efforts and collaboration.
A wide range of information retrieval strategies was tested in different TREC experiments (i.e. from TREC 1 in 1992 to TREC 16 in 2007). Some notable examples are:

- Boolean search;
- Statistical and probabilistic indexing and term weighting strategies;
- Passage or paragraph retrieval;
- Combining the results of more than one search;
- Retrieval based on prior relevance assessments;
- Natural language-based and statistically based phrase indexing;
- Query expansion and query reduction;
- String and concept-based searching;
- Dictionary-based searching;
- Question-answering;
- Content-based multimedia retrieval; and
- Relevance judgements.

**Self Check Exercise**

**Note:**

i) Write your answers in the space given below.

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

8) What do you mean by the evaluation of an indexing system? What are its purposes?

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

9) How do you measure the ‘efficiency’ and ‘effectiveness’ of an indexing system?

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

10) How recall and precision ratios help in measuring the performance of an indexing system?

........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
11) Distinguish between ‘Relevance’ and ‘Pertinence’.

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

12) What is the significance of TREC?

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

9.7 SUMMARY

In this Unit we have dealt with a comprehensive view of the different issues and perspectives of subject indexing. It begins with a brief overview of the origin and development of subject indexing from the ancient to modern times. Then, meaning and purposes of subject indexing have been discussed. It is impossible to understand the subject indexing properly without being aware of the different principles and processes associated with it. For this, different aspects of indexing principles and processes are discussed. Indexing policy with particular reference to exhaustivity and specificity has been explained. Problems associated with subject indexing are highlighted. The quality of indexing is defined in terms of consistency in indexing. This Unit lays major emphasis on the evaluation of indexing system forming an essential component of the information and retrieval system. Different evaluation criteria like recall, precision and relevance are discussed. Major retrieval experiments are discussed briefly at the end of this Unit.

9.8 ANSWERS TO SELF CHECK EXERCISES

1) Subject indexing is a set of interrelated processes associated with the representation of informational content of the document in the form of a subject index. A subject index is a systematic guide to the text of any reading matter or to the contents of documents, comprising a series of entries, with headings arranged in alphabetical or other chosen order and with references to show where each item indexed is located. It is a working tool designed to help the user to find his required documents in a given subject field. It gives subject access to documents irrespective their physical forms.

2) The purposes of subject indexing are to: satisfy subject approach to information; identify pertinent materials on a given subject or topic; enable the enquirer to find materials on related subjects through the network of references; prescribe a standard methodology for constructing uniform subject headings; bring consistency in the choice and rendering of subject entries, using standard vocabulary and according to the given rules and procedures; enable users in accessing any desired document(s) from the catalogue or index; economize the bulk and cost of indexing by deciding on the optimum number of index entries; and provide user-friendly approach in naming the subjects through any vocabulary common to a considerable group of users—specialists or laymen.
The origin and development of subject indexing languages is intimately related with the historical development of libraries through ancient and medieval periods to modern days. The specific usage of the term index goes back to ancient Rome when term index was used for the little slip attached to papyrus scrolls on which the title and sometimes name of the author of the work was written in order to identify the scrolls on the shelves. In the first century A.D., the meaning of the word was extended from “title” to a table of contents or a list of chapters and hence to a bibliographical list or catalogue. During the 13th century alphabetisation by names of authors under systematic subject arrangement was well known with the invention of printing. The practice of supplementing systematic listing with an alphabetical subject index was introduced with the development of university libraries during 15th century. With the rapid growth of knowledge in many fields in the course of the 19th century and the resulting increase in the volume of books and periodicals, subject indexing got a new momentum. In the 19th century, subject access to books was by means of a classification. But it could not go deep enough to the extent of individualizing subjects of documents. Only in the late 19th century, alphabetical subject indexing became widespread and more systematic. It was Charles Ammi Cutter who first gave a generalized set of rules for subject indexing in his Rules for a Dictionary Catalogue (RDC) published in 1876. Cutter’s rules for specific subject headings for use in a dictionary catalogue seemed to have made an impact on the development of readymade lists of subject headings like Library of Congress Subject headings and Sears List of Subject Heads in 1897 and 1923 respectively. The first quest for logical approach towards solving the problems of compound subject headings is evident in J. Kaiser’s Systematic Indexing (1911). Kaiser was the first person who gave the idea of categorization of terms: concrete and process. Dr. S. R. Ranganathan was the first to analyse the universe of subjects in depth and developed chain indexing system to supplement the classified catalogue in 1934. Qualities of the classification scheme determined the qualities of the subject headings drawn according to chain procedure in view of its total dependency on the classification scheme. J. E. L. Farradane proposed a new line of thinking in subject indexing system by suggesting the relationship between pairs of concepts instead of categorisation of concepts in Relational Indexing developed in the early 1950s. The contribution of E. J. Coates in subject indexing (1960) was not original in nature. Coates merely synthesised the ideas of Cutter, Kaiser, Ranganathan and Farradane. Preserved Context Index System (PRECIS), developed by Derek Austin in 1969 and applied in BNB in 1971 as an alternative to the chain procedure is considered as the first computerised pre-coordinate indexing system. Postulate-based Permutted Subject Indexing (POPSI) developed in 1969 sought to rectify the defects of chain procedure from entirely different perspective. It is evident from the above discussion that the research on development and use of various subject indexing systems was devoted to techniques of constructing pre-coordinate subject heading. Difficulties stemming from the question pre-coordination of terms led to the development of post-coordinate indexing or simply coordinate indexing systems like Uniterm, Optical Coincidence Card / Peek-a-boo, Edge-Notched Card, etc during 1960s. Computers began to be used to aid information retrieval system in the 1950s. H P Luhn has been credited for the production of computerized permuted title indexes, which he named as Keyword-In-Context (KWIC) index. The publication of Science Citation Index (SCI) by Eugene Garfield in 1963 provided a new approach to the bibliographic file organization. In the arena of computerized indexing, there has been considerable research on the user-interface design, indexing systems using Artificial Intelligence
techniques like Natural Language Processing (NLP), Knowledge Representation Model and Expert System-based subject indexing systems. As a result of the phenomenal growth of content on the Web as an indexing problem, we have seen a continued interest in the development of tools and techniques to index the Web resources. Different search tools and technologies were developed in finding the resources on the Web so far to make computers understand the semantics underlying contents of the Web resources.

4) The indexing process basically consists of two intellectual steps: Conceptual analysis and Translation. Conceptual analysis involves the identification of different component ideas associated with the thought content of the document and establishment of interrelationship between those component ideas. It is carried out by examining the document for establishing its subject content and identifying the principal concepts present in the subject of the document. Translation refers to the expression of principal concepts as identified while analysing the thought content of the document into the language of the given indexing system.

5) The indexing principles operate at four levels: user as the focus, utility, common usage, and specificity.

6) Exhaustivity is the measure of the extent to which all the distinct subjects are discussed in a particular document are recognised in indexing operation. Exhaustivity in indexing allows for the recognition of concepts embodied not only in the main theme of the document but also in sub-themes of varying importance. 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. 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.

7) An indexing failure may take place at the following stages of indexing process: (a) failure in establishing concepts during conceptual analysis of the content of a document; (b) failure to identify a topic that is of potential interest to the target user group; (c) misinterpretation of the content of the document, leading to the selection of inappropriate term(s) (d) failure in translating the result of conceptual analysis into the indexing language; (e) failure to use the most specific term(s) to represent the subject of the document; (f) 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 (g) omission of important term(s).

8) By evaluation of an Indexing system we mean the measurement of the performance of the indexing system—success or failure, in terms of its retrieval efficiency (ease of approach, speed and accuracy) to the users, and its internal operating efficiency, cost effectiveness and cost benefit to the managers of the system. The purposes of the evaluation of indexing are to (a) identify the level of performance of the given indexing system, (b) understand how well the given indexing system fulfills the queries of the users in retrieving the relevant documents, (c) compare the performance of two or more indexing systems against a standard, (d) identify the possible sources of failure of the given indexing system or inefficiency with a view to raising the level of performance at some future date, (e) justify the existence of the given indexing system by analysing its costs and benefits, (f) To establish a
foundation of further research on the reasons for the relative success of alternative techniques, and (g) improve the means employed for attaining objectives or to redefine goals in view of research findings.

9) The effectiveness of an indexing system can be measured by calculation of recall and precision ratios. Efficiency can be measured by what minimum cost and effort does the indexing system function effectively. Cost factors are calculated indirectly, such as response time (i.e. that is time taken by the system to retrieve the information), user effort (i.e. the amount of time and effort required by a user to interact with the indexing system and analyse the output retrieved in order to get the required information), the cost involved.

10) Recall refers to the index’s ability to let relevant documents through the filter. Recall ratio is a ratio of the relevant documents retrieved to the total number of relevant documents potentially available. It measures the completeness of the output by means of a ratio called recall ratio as mentioned below:

Recall ratio = \( \frac{R}{C} \times 100 \)

Where, \( R \) = Total Number of relevant documents retrieved against a search

\( C \) = Total number of relevant documents to that particular request in the collection.

When the system retrieves items that are relevant to a given query it also retrieves some documents that are not relevant. These non-relevant items affect the success of the system because they must be discarded by the user, which results the wastage of significant amount of time. The term ‘Precision’ refers to the index’s ability to hold back documents not relevant to the user. Precision ratio is a ratio of the relevant documents retrieved to the number of documents retrieved. The formula for calculation of Precision Ratio is:

Precision Ratio = \( \frac{R}{L} \times 100 \)

Where, \( R \) = Total number of relevant documents retrieved against a search

\( L \) = Total number of documents retrieved in that search

11) The term pertinence refers to a relationship between a document and an information need, whereas the term relevance refers to a relationship between a document and a request statement (i.e. expressed information need). Relevance is consensus judgment, but pertinence relates to an individual judgment.

12) Research in information retrieval was limited to small test collections prior to 1991 and there was little transfer of research ideas into commercial systems. The Text REtrieval Conference (TREC), launched in 1991, introduced the first large test collection of full-text documents in order to enable IR researchers to scale up from small test collection of data to larger experiments on a realistic scale. TREC is an open testing event conducted annually with common tasks and a standard evaluation scenario. Every year, a set of information retrieval/search tasks is defined. The TREC series of experiments has drawn attention of the LIS professionals all over the world since its inception and has shown that significant research results can be obtained through international efforts and collaboration. A wide range of information retrieval strategies was tested in different TREC experiments (i.e. from TREC 1 in 1992 to TREC 16 in 2007).
### 9.9 KEYWORDS

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class number</td>
<td>Notation that designates the class to which a given item belongs.</td>
</tr>
<tr>
<td>Concept</td>
<td>A unit of thought, formed by mentally combining some or all of the characteristics of a concrete or abstract, real or imaginary object. Concepts exist in the mind as abstract entities independent of terms used to express them.</td>
</tr>
<tr>
<td>Conceptual Analysis</td>
<td>A process which involves deciding what a document is about—that is, identification of different component ideas associated with the thought content of the document and establishment of interrelationship between those component ideas.</td>
</tr>
<tr>
<td>Exhaustivity</td>
<td>The use of enough terms to cover the all topics discussed in a document. It relates to the breadth of coverage in indexing. It is sometimes called depth indexing.</td>
</tr>
<tr>
<td>Identifier</td>
<td>A significant term, number, acronym, name, or symbol used alone or with other identifiers to refer to a library resource.</td>
</tr>
<tr>
<td>Index</td>
<td>A detailed alphabetical or numerical list. List entries represent an aspect of a bibliographic record and are organised into searchable files used to retrieve records in a database or set of records.</td>
</tr>
<tr>
<td>Indexing</td>
<td>The process of evaluating information entities and creating indexing terms, normally subject or topical terms that aid in finding and accessing the entity. Index terms may be in natural language or controlled vocabulary or a classification notation.</td>
</tr>
<tr>
<td>Indexing Term</td>
<td>The representation of a concept in an indexing language, generally in the form of a noun or noun phrase. Terms, subject headings, and heading-subheading combinations are examples of indexing terms. Also called Descriptor.</td>
</tr>
<tr>
<td>Precision Ratio</td>
<td>The term precision relates to the ability of an indexing system not to retrieve irrelevant items. Precision ratio refers to the proportion of retrieved items that are relevant.</td>
</tr>
<tr>
<td>Quality of Indexing</td>
<td>The ability to retrieve what is wanted and to avoid what is not wanted.</td>
</tr>
<tr>
<td>Recall Ratio</td>
<td>The term recall refers to a measure of whether or not a particular item is retrieved or the extent to which the retrieval of wanted items occurs.</td>
</tr>
</tbody>
</table>
Recall ratio is the proportion of relevant items retrieved and thus, it is a measure of the completeness of a search in an index file.

Relative Index: A type of index in which various aspects of an idea and its relationship with other ideas are shown.

Relevance: Relevance refers to the ability of an information retrieval system to retrieve material that satisfies the information needs of the users.

Specific index: Specific index is an index which shows the broad topics in the form of one-idea-one-entry, i.e. specific context of a specific idea.

Specificity: It refers to the use of much smaller number of terms to cover only the central subject matter of a document. The more specific the terms used, the fewer the entries per term on the average. Specificity is the property of the vocabulary used in indexing and it relates to the depth of treatment of the content of a document in indexing.

Subject: A systematised or organised body of ideas.

Subject cataloguing: The part of cataloguing that provides subject heading/terms and/or classification.

Subject Headings: An alphabetical list of words or phrases that represent a concept that is under authority control, e.g., the Library of Congress Subject Headings.

9.10 REFERENCES AND FURTHER READING


