UNIT 11 INDEXING TECHNIQUES

Structure

11.0 OBJECTIVES

After reading this Unit, you will be able to:

- understand the basic principles of subject indexing techniques;
- appreciate the differences between (a) derived and assigned indexing, (b) pre- and post-coordinate indexing systems;

- trace the major contributions in pre-coordinate indexing systems;
- learn the different stages of intellectual operations and working of some important pre-coordinate indexing systems such as, Chain Indexing, PRECIS and POPSI;
- understand the objective conditions that led to the development of post-coordinate indexing;
- learn the different entry structure such as item entry as well as term entry systems and the methods of post-coordinate indexing with particular reference to Uniterm indexing system and post-coordinate searching devices;
- explain and apply the different varieties of keyword indexing;
- appreciate the differences between the manual indexing and computerised indexing;
- understand the concept of computerised indexing in terms of its features, components, categories, and index file organisation;
- learn different methods associated with the generation of index entries with the aid of computer;
- get acquainted with the indexing internet resources with particular reference to search engine indexing and other associated concepts; and
- develop skills of using different indexing techniques for formulating different types of subject headings.

11.1 INTRODUCTION

Subject approach to information has been an area of intense study and research in the area of organisation of information, resulting in the generation of new theories and the design of the corresponding new indexing techniques based on these theories. Indexing technique actually originated from what is known as the 'back-of-the-book index. Its objective is to show where exactly in the text of a document a particular concept (denoted by a term) is mentioned, referred to, defined or discussed. The 'back-of-the-book index' may be either in the form of specific index or relative index. Specific index presents the broad topics in the form of one-to-one-entry whereas the relative index is one which displays each concept in different context. The best example of such an index is the relative index of Dewey Decimal Classification. But the relative index is usually unique to the text to which it points to and is quite difficult to maintain on a large scale. Subsequently we have seen the development of pre-coordinate indexing model.

Till about the early fifties of the last century, the pre-coordinate indexing models were the only ones that had been developed. In the subsequent decades, the post-coordinate indexing models were designed and developed, the physical apparatus for these index files also changed from the conventional index cards to different formats of post-coordinate indexing models. With the advent of the computers, the keyword index models like KWIC, KWOC, and KWAC were introduced, post-coordinate indexing also became more amenable for computer manipulation. Most of the bibliographic databases today have indexes based on post-coordinate indexing principles. The following sections of this Unit present the major developments in subject indexing techniques for organising the index file.

11.2 DERIVATIVE INDEXING AND ASSIGNMENT INDEXING

Indexing can be either 'derived indexing'/'derivative indexing' or 'assigned indexing'/ 'assignment indexing'. Derived indexing is a method of indexing in which a human indexer

or computer extracts from the title and/or text of a document one or more words or phrases to represent the subject(s) of the work, for use as headings under which entries are made. It is also known as *extractive indexing*.

In derivative indexing, terms to be used to represent the content of the document are derived directly from the document itself. Here no attempt is made to use an indexer's own knowledge of the subject or other guides, but use only the information which is manifest in the document. Index terms are derived directly from the title or text of the document. It requires least intellectual effort on the part of the indexer. Mechanical devices and computers are used in abundance to carry out the burden of index preparation as well as the tasks of matching these with the questions of users put to the system. Examples of derivative indexing are keyword indexing, citation indexing, automatic indexing, etc.

Assigned indexing is also known as 'concept indexing', because it involves identifying concept(s) associated with the content of each document. It is a method of indexing in which a human indexer selects one or more subject headings or descriptors from a list of controlled vocabulary to represent the subject(s) of a work. The indexing terms selected to represent the content need not appear in the title or text of the document indexed. Here, an indexing language is designed and it is used for both indexing and searching. Some notable examples of assignment indexing are chain indexing, PRECIS, POPSI, classification schemes, etc.

11.3 PRE-COORDINATE INDEXING SYSTEMS

Basically all indexing systems are, by nature, coordinate indexing. The purpose of using a combination or coordination of component terms is to describe the contents of the documents more precisely. Many subjects can be expressed in a single term, e.g. Libraries, Cataloguing, Management, etc. Others are expressed as a combination of these, e.g. Cataloguing in libraries, Management of libraries, etc. When the indexer assigns subject headings representing such compounds and arranges entries in a series of classes according to the subject content of the document, the resulting system is referred to as *pre-coordinate indexing system.* Here, the indexer coordinates the component terms representing compounds at the input stage, i.e. at the time of indexing in anticipation of users' approach. Most of the classification schemes allow a measure of coordination, either by including compound subjects or by providing facilities for creating them out of simple elements. The same is true for readymade lists of subject headings like Sears List of Subject Headings and Library of Congress Subject Headings. Such classification schemes and lists of subject headings can therefore be regarded as pre-coordinate indexing languages. Almost all pre-coordinate indexing models have an *a priori* approach in choosing their semantic paradigms such as categorisation of concepts, role indicators, relational operators, etc. Pre-coordinate indexing involves coordinating (combining, pulling together concepts) followed by engaging in an act of *synthesis* to build the index entries. In such a system, the most important aspect is to determine the order of significance by following the syntactical rules of the given indexing language.

11.3.1 Cutter's Contribution

Charles Ammi Cutter first set forth the rules for alphabetical subject headings in a systematic way in his *Rules for a Printed Dictionary Catalog* in 1876. He used the term 'subject cataloguing' instead of 'subject indexing'. It was C. A. Cutter who gave the idea of specific subject entry. He also tried to systematise the rules for compound subject headings which consisted of more than one term as a phrase and where the

subject heading was composed of a name of a subject and the name of a locality and so on. Cutter's concept of specific subject heading was quite different from what we mean today. He had in mind a set of stock subject names and every book had to be accommodated under the most restricted subject which contained the subject of a book. For compound subject heading, Cutter laid down the rule that the order of the component terms in compound subject heading should be the one that is decidedly more significant. But Cutter could not prescribe how one will come forward to decide which one is more significant. The question of significance varies from user to user. The decision in respect of 'significance' was left to the judgment of individual indexer, which was subjective. Some specific rules/guidelines as furnished by Cutter in his *Rules for a Printed Dictionary Catalog* are mentioned below:

1) **Person versus Country**

Entry will be under person in case of single / personal biography. Entry will be under country in case of history, event, etc. For example the biography of Sachin Tendulkar would entered under his name whereas the biography of Indira Gandhi would be entered her name and India- history also.

2) **Country versus Event**

Entry of an event would be under event if it is proper noun, e.g. Jalian Wala Bagh. Entry would be under country if it is common noun. e.g. Freedom fighters in India, entry would be under India.

3) **Subject versus Country**

In scientific subjects, entry will be under subject qualified by place. e.g. Oceanography, India.

It has been stipulated to prepare entry under the name of place qualified by the subject in subject areas like History, Government and Commerce, e.g. India—Moghul Period.

For Humanities, Literature, Art, etc., adjectival form of subject headings were suggested, e.g. Indian Painting, Moghul Architecture, etc.

4) Between overlapping subjects entry will be according to the importance of the subjects. It is to be pointed out here that the decision regarding 'importance' was left to the judgment of the indexer.

5) When there is a choice between different names, Cutter prescribed the followings:

   a) Language—If there are two languages out of which one is English, entry will be under English. Here the rules appear to be biased towards English language.

   b) Synonyms—Entry will be under one word with reference to others.

   c) Antonyms—Entry will be under one word with reference to others.

6) In case of compound subject headings, Cutter prescribed the following rules:

   a) A noun preceded by an adjective, e.g. Organic Chemistry, Ancient History, etc.

   b) A noun preceded by another noun used as an adjective, e.g. War Prisoners, Flower Fertilisation, Death Penalty, etc.

c) Use direct form in case of a noun connected to another noun with a preposition, e.g. Patient with heart disease, Fertilisation of flowers, death penalty, etc.

d) Use direct form in case of phrase or sentence used as the name of a subject, e.g. Medicine as profession.

11.3.2 Kaiser's Contribution

Kaiser started from the point where Cutter left. In 1911, Julius Otto Kaiser (1868–1927), a special librarian and indexer of technical literature, developed a method of subject indexing system known as *Systematic Indexin*g. Kaiser defined indexing as "the process by which our information is collected and made accessible" and claimed that it constitutes "the main work of organising information". The definition highlights a cardinal tenet of his theory of systematic indexing —that, within the framework of a business library with which he was attached, the primary aim should be not to index 'documents' but 'information', which Kaiser took to be the various 'facts and opinions' (i.e., informational units) encoded in the texts of documents. He viewed systematic indexing as a two-step procedure: The first step was to analyse a subject so as to distinguish constituent concepts associated with the content of the given document into two fundamental categories of facets: "Concretes", and "Processes". 'Concrete' refers to things, places and abstract terms, not signifying any action or process; e.g. Gold, India, Physics, etc. "Process" refers to mode of treatment of the subject by the author (e.g. *Evaluation* of IR system, *Critical analysis* of a drama), an action or process described in the document (e.g. *Indexing* of web documents), and an adjective related to the concrete as component of the subject (e.g. Strength of a metal).

Kaiser also laid a rule that if a subject dealing with place, double entry (*'Concrete—Place—Process'* and *'Place—Concrete—Process'*) is to be made. For example, index entries for 'Manufacturing of petrochemicals in West Bengal' would be: PETROCHEMICALS—*West Bengal—Manufacturing; WEST BENGAL—Petrochemicals— Manufacturing*. The second step was to synthesize constituent concepts into indexing statements formulated according to the strict rules of citation order—Concrete is to be followed by Process.

Thus, according to Kaiser's systemic indexing, all indexing terms should be divided into fundamental categories "concretes", "countries", and "processes", which are then to be synthesized into indexing "statements" formulated according to strict rules of citation order. Some examples of subject headings according to Kaiser's systematic indexing are furnished below:

Documents	Categories	Subject Headings
Indexing of films	Concrete—Process	Films—Indexing
Strike in India	Country—Process	India—Strike
Libraries in Nepal	Concrete—Country	Libraries—Nepal
Manufacturing of copper in Assam	Concrete—Country—Process	Copper—Manufacturing—Assam

Kaiser's systematic indexing did not make any provision for entry under the 'process' term and as a result it failed to satisfy the users' approach by the 'process' term. The concept of 'time' was also left by Kaiser. It is to be pointed out here that Kaiser was perhaps the first person who gave the idea of categorisation in subject indexing.

11.3.3 Chain Indexing

Ranganathan's facet analysis of subject provides a kind of representation of subjects by transforming multidimensional relations of subject into a modulated layer of linear representation. Ranganathan is credited with the invention of chain indexing, an economical system of providing access to the terms in classification schedules without replicating the hierarchical structure of the classification in the alphabetical index. Ranganathan's Chain indexing technique was devised as a complementary and supplementary tool to classification schemes. However, due to the efficiency and economy, this technique can effectively be made use of in deriving alphabetical subject indexes for any indexing/abstracting services.

Definition and Use

A chain is a string of terms organised in a particular sequence based on the classification scheme that the chain adopts. The sequence of terms is pre-coordinated according to the syntactical rules of the given classification scheme. Chain indexing is a method of deriving subject headings from the chain of successive subdivisions leading from general to specific level needed to be indexed.

According to Ranganathan, chain indexing is a

"*procedure for deriving class index entry (i.e. subject index entry) which refers from a class to its class number in a more or less mechanical way.*"

A note is also given with the above definition as

"*Chain procedure is used to derive class index entries in a Classified Catalogue, and specific subject entries, subject analytical, and 'see also' subject entries in a Dictionary Catalogue*."

Chain indexing was used in *the British National Bibliography* (BNB) in the 1950s and 1960s until it was replaced by PRECIS-indexing.

Chains and Links

The concept of 'chain' is the foundation of chain indexing. A chain is deemed to be a structural manifestation of a subject. The term 'structure' in this context refers to the parts constituting a subject and their mutual interrelationship. It is a modulated sequence of sub-classes or isolates ideas.

Since the chain expressed the modulated sequence of sub-classes more effectively in a notational scheme of classification of subjects, this method takes the class number of the document concerned as the base for deriving subject headings not only for specific subject entry but also for subject reference entries. The nature and structure of the classification scheme used to classify the subject of the document controls the structure of the subject headings drawn according to the chain procedure. The concept of 'chain' becomes operative only after the concept of a set 'links' about the structure of the subject is conceded. A chain should comprise a link of every order that lies between the first link and last link of the chain. The different types of links in chain indexing system are discussed below:

- **Sought Links (SL):** Sought links denote the concepts (at any given stage of the chain) that the user is likely to use as access points.
- **Unsought Links (USL):** Unsought links denote those concepts that are not likely to be used as access points by the user.

- **False Links (FL):** False links are those that really do not represent any valid concept, mostly these are connecting symbols or indicator digits.
- **Missing Links (ML):** Missing links represent those concepts that are not available in the preferred classification scheme, these are inserted by the indexer by means of verbal extension at the chain-with-gap corresponding to the missing isolate in the chain whenever there is such a need.

Steps in Chain Indexing

The following steps are to be followed in chain indexing for deriving different types of subject headings:

1) **Construction of the Class Number of the Subject of the Document**

Classify the subject of the document by following a preferred classification scheme. A class number constructed according to a scheme for notational classification will form the basis for applying the rules for chain procedure for deriving subject headings.

2) **Representation of the Class Number in the form of a Chain**

Represent the class number in the form of a chain in which each link consists of two parts: class number and its verbal translation in standard term or phrase used in the preferred classification scheme.

3) **Determination of Links**

Determine different kinds of links: Sought Links (SL), Unsought Links (USL), False Links (FL), and Missing Links (ML).

4) **Preparation of Specific Subject Heading**

Derive specific subject heading for the specific subject entry from the last sought link and moving upwards by taking the necessary and sufficient sought links in a reverse rendering or backward rendering process. If the subject includes a space isolate, time isolate or a form isolate, break the chain into different part(s) at the point(s) denoting space, time and form in the class number. In such a situation, specific subject heading is to be derived from last SL of first part in reverse rendering process and then by second part, third part, etc., if any, in the similar process. Then, the components of derived subject headings are to be arranged in the sequence of their derivation from each part of the chain of the class number.

5) **Preparation of Subject Reference Headings**

Derive subject reference heading for the subject reference from each of the upper sought links. This process continues until all the terms of upper sought links are exhausted and indexed.

6) **Preparation of Subject Reference Entries**

Prepare subject reference entries or 'see also' references from each subject reference heading to its specific subject heading. When a subject heading starts from last sought link denoting space or time or form, prepare 'See' references instead of 'See also' references from subject reference heading(s) to specific subject heading.

7) **Preparation of Cross References, if any**

Prepare cross references (i.e. 'see' references) for each alternative and synonymous term/heading used in the specific as well as subject reference headings.

8) **Alphabetisation**

Merge specific subject entries, subject references (i.e. 'see also' references) and 'see' references and arrange them in single alphabetical sequence.

The above noted steps in chain indexing are demonstrated below with illustrative example:

0) **Subject of the Document**

Researches on Child Psychology in India

1) **Class Number of the Subject of the Document**

Class no.: 155.4072054 [according to DDC, 22nd Edition]

2) **Representation of the Class Number in the form of a Chain**

100	Philosophy, Parapsychology and Occultism
150	Psychology
155	Differential and developmental psychology
155.	———
155.4	Child psychology
152.40	———
152.407	Education, Research related topics
152.4072	Research
152.40720	———
152.407205	Asia
152.4072054	India

3) **Determination of Links**

100	Philosophy, Parapsychology and Occultism [USL]
150	Psychology [SL]
155	Differential and developmental psychology [USL]
155.	[FL]
155.4	Child psychology [SL]
152.40	[FL]
152.407	Education, Research related topics [USL]
152.4072	Research [SL]
152.40720	[FL]
152.407205	Asia [USL]
152.4072054	India [SL]

4) **Preparation of Specific Subject Heading**

Research, Child psychology, India

5) **Preparation of Subject Reference Headings**

Research, Child psychology

Child psychology

Psychology

6) **Preparation of Cross References**

India, Research, Child psychology

7) **Preparation of Index Entries**

Research, Child psychology, India 152.4072054

Bibliographical description and abstracts of the document are to be furnished under the specific subject heading.

Research, Child psychology 152.4072

See also

Research, Child psychology, India

Child Psychology 155.4

See also

Research, Child psychology, India

Psychology 150

See also

Research, Child psychology, India

India, Research, Child psychology

See

Research, Child psychology, India

8) **Alphabetisation**

Arrange the above entries according to single alphabetical order.

Advantages

A major advantage of chain indexing is that it ensures the collocation of aspects of a subject which have been scattered in the classification scheme (i.e. distributed relatives) because last link in the class number is always the first link in the chain of subject index entries.

It offers general as well as specific information to all information seekers by deriving subject headings from the chain of successive subdivisions that leads from the general to most specific level.

It is more or less a mechanical system and it is economic also.

Criticisms

- Too much dependency on classification schemes

Chain indexing could be operative effectively only if they are backed up by a good structured classification scheme. If chain indexing is based on a structurally defective classification scheme, the subject headings drawn according to chain procedure will naturally become defective.

- Disappearance of Chain

Chain disappears in each stage of deriving subject reference entries and thus it results in the loss of full context of the content of the document.

- **Lack of Specificity**

Chain indexing provides only one specific entry and others are subject references.

- **Unsuitability for computerisation**

The formation of 'Chain' is very much a human intellectual operations which is beyond the capability of the computer to manipulate.

- **Running from pillar to post**

It makes direct access to specific subject heading possible only for those who knew the specific subject heading for the given document, but in most cases access to specific subject heading was possible at the cost of running from pillar to post since only one entry is specific subject entry and others are cross references in chain indexing system.

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) Distinguish between Derived indexing and Assigned indexing.

.....

.....

.....

.....

2) Discuss the techniques by which J. Kaiser has solved the problems of indexing compound subject.

.....

.....

.....

.....

3) What is the epithet of the term 'chain' in chain indexing? Why is the class number of the subject taken as the base for deriving subject headings according to chain indexing system?

.....

.....

.....

.....

4) Enumerate the steps involved in chain indexing.

.....

.....

.....

.....

11.3.4 PRECIS (Preserved Context Index System)

PRECIS is generally considered as one of the most ambitious late twentieth century attempts to create an indexing system from scratch. A major break-through in the field of subject indexing was achieved by Classification Research Group (CRG), London and Derek Austin came out with a new method of subject indexing called PREserved Context Index System (PRECIS) in the early 1970s for the *British National Bibliography (BNB)*. When the BNB began in 1950, it used chain indexing system for about 20 years. However it was not, for various reasons, ideal for a computerised system, and in 1971, when BNB had developed the MARC system in the United Kingdom and was also engaged in using computers for the production of BNB itself, chain indexing was replaced by PRECIS.

What is PRECIS?

PRECIS is a system of subject indexing in which the initial string of terms organised according to the scheme of role operators, is computer manipulated in such a way that each sought term in the string functions as the approach term while preserving the full context of the document. Entries are restructured at every step in such a way that the user can determine from the format of the entry which terms set the approach term into its context and which terms are context dependent on the approach term.

Objectives of PRECIS

a) The computer, not the indexer, should produce all index entries. The indexer's responsibility is to prepare the input strings and to give necessary instructions to the computers to generate index entries according to definite formats.

b) Each of the sought terms should find index entries and each entry should express the complete thought content / full context of the document unlike the chain procedure where only one entry is specific—i.e. fully co-extensive with the subject of the document and others are cross references describing only one aspect of the thought content of the document.

c) Each of the entry should be expressive.

d) The system should be based on a single set of logical rules to make it consistent.

e) The system should be based on the concept of open-ended vocabulary, which means that terms can be admitted into the index at any time, as soon as they have been encountered in the literature.

f) The system must have sufficient references between semantically related terms.

Features of PRECIS

- It is more amenable to automatic manipulation than indexing based on the notational classifications.
- The permuted entries read naturally, which is achieved by the prescribed order of the role operators;
- The terms are linked to a machine-held thesaurus thereby providing possible '*see'* and '*see also'* references*;*
- PRECIS can be adapted to other languages.
- The indexer determines the meaning of the terms codes the roles and identifies the lead terms, whereas the computer takes care of the permutations.
- Its subject formulation is completely independent of classification, therefore exclusively geared to no classification numbers assigned in the MARC record.
- Context is preserved: It presents the full subject statement at every point of index entry, by gradual inversion of the concept string, thus overcoming the problem of the disappearing chain.

Principles of PRECIS

Two principles are followed in PRECIS:

a) **Principle of Context Dependency**: The "context-dependency" principle may be seen as a combination of context and dependency. When this principle is followed in a PRECIS input string, each term is qualified and sets the next term into its wider context. In other words, the meaning of each term in the string depends upon the meaning of its preceding term and taken together, they all represent the single context. Each term is hence dependent, directly or indirectly, on all the terms which precede it.

b) **Principle of One-to-One Relationship**: When the terms are organised according to the principle of context dependency, they form a one-to-one related sequence—each of the terms in the string directly related to its next term.

Syntax and Semantics

The syntax of PRECIS is based on the role operators, codes and logical rules which act as instruction to the computer. The semantics of PRECIS is handled by linking the terms to a machine-held thesaurus thereby providing possible '*see'* and '*see also'* references.

Role Operator: Role operators consist of a set of alpha-numeric notations which specifies the grammatical role or the function of the indexed term and regulates the order of terms in the input string. Role operators and their associated rules also serve as the computer instruction for determining the format, typography and punctuation associated with each index entry. There are two kinds of role operators: Primary operators and secondary operators. Primary operators control the sequence of terms in input string and determine the format of index entries. Any of the secondary operators is always to be preceded by the primary operator to which it relates.

Codes: Use of codes in the string brings expressiveness in the resulting index entries. Three types of codes are there: Primary, Secondary and Typographic codes.

Input String: A set of terms arranged according to the role operators which act as instructions to the computer for generating index entries.

Schema of Role Operators

Primary Operators		
Environment of core concepts	0	Location
Core concepts	1	Key System *Thing when action not present. Thing towards which an action is directed, e.g. object of transitive action, performer of intransitive action*
	2	Action; Effect of action
	3	Performer of transitive action (Agent, Instrument); Intake; Factor
Extra-core concepts	4	Viewpoint-as-form
	5	Selected Instance: study region, study example, sample population
	6	Form of document; Target user
Secondary Operations		
Co-ordinate concepts	f	'Bound' co-ordinate concept
	g	Standard co-ordinate concept
Dependent elements	p	Part; Property
	q	Member of quasi-generic group
	r	Assembly
Special classes of action	s	Roll definer; Directional property
	t	Author-attributed action
	u	Two-way interaction

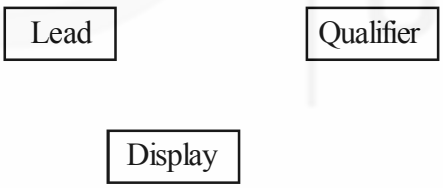
Schema of Codes

Primary codes	
Theme Interlinks	$x 1st concept in coordinate theme
	$y 2nd/subsequent concept in co-ordinate theme
	$z Common concept
Term Codes	$a Common noun
	$c Proper name (class of-one)
	$d Place name

Secondary codes
Differences
Preceding differences (3 characters)— 1st and 2nd characters: $0 Non-lead, space generating $1 Non-lead, close-up $2 Lead, space generating $3 Lead, close-up 3rd character: Number in the range, 1 to 9 indicating level of difference
Date as a difference $d
Parenthetical difference $n Non-lead parenthetical difference $o Lead parenthetical difference
Connectives $v Downward reading connective $w Upward reading connective
Typographic codes
$e Non-filing part in italic preceded by comma
$f Filing part in italic preceded by comma
$g Filing part in roman , no preceding punctuation
$h Filing part in italic preceded by full point
$I Filing part in italic , no preceding punctuation

Entry Structure of PRECIS

The entry structure of PRECIS string consists of a two-dimensional display rather than the one dimensional that we have been accustomed to; instead of putting everything on one line, so that the only relationship which could be shown was that of following or proceeding, PRECIS uses two-line-three part entry structure as follows:



Lead is occupied by the approach term, which is the filing word and is offered as the user's access point in the index.

Qualifier position is occupied by the term(s) that sets the lead into wider context (i.e. general to specific). Together, the Lead and the Qualifiers correspond to the Heading. Terms in the heading set down in a narrower-to-wider context order. When the first term of the input string appears in the Lead position, the Qualifier position is usually kept blank.

Display position is occupied by those additional set of qualifying terms of the PRECIS string, which rely upon the heading for their context. When the last term of the input string appears in the Lead position, the Display position becomes empty.

Steps in PRECIS

Let us take the following title of a document for demonstrating the different steps of subject indexing according to PRECIS:

0. **Title**: University libraries in West Bengal

1. **Analysis**: Involves analysis of the thought content of the document and formulating the subject statement in natural language:

Measurement of the performance of university libraries in West Bengal

2. **Preparation of Input String**: Involves the identification of the status or role of each component term denoting key concept in terms of the role operators of PRECIS and assigning the appropriate operators to prepare an input string. The stages of the preparation of input string are furnished below:

[For understanding the stages of the preparation of input string, students are required to consult the schema of role operators and codes as furnished under the sub-section 11.3.4. of this Unit]

a) Identifying the concept signifying an action (if there be any).

In the present example, the action concept is denoted by the term 'Measurement' and this term should, therefore be prefixed by the role operator (2).

b) Identifying the kind of action represented by this term, i.e. whether transitive action or intransitive action. 'Measurement' is clearly a transitive action since it is capable of taking an object. The object of transitive action is considered as the key system and is coded by the operator (1). In the present example, it is the 'Performance' which is being measured, so the input string should now appear in the form:

(1) performance

c) Identifying the concept, if any related as property and/or whole-to-part. In this example, 'performance' is the property of 'university libraries' and 'libraries' is the part of the 'universities'. As a result, we will get the following input string:

(1) universities

(p) libraries

(p) performance

(2) measurement

d) The remaining term 'West Bengal' signifies the environment (i.e. geographical location) in which the whole thing takes place. Accordingly, the operator (0) is to be prefixed to the concept and the resulting input string will be:

(0) "West Bengal

(1) "universities

(p) "libraries

(p) "performance

(2) "measurement

Note: The primary operators have ordinal filing values and the terms in the above input string are sequenced accordingly. Thus, component terms are organised into the above input string according to the principle of context dependency. The secondary operator (p), prefixed with 'libraries' and denoting 'part' of the 'universities', is preceded by the primary operator (1) to which it is related. Similarly,

the secondary operator (p), prefixed with 'performance' and also denoting 'property' of the 'libraries', is preceded by the secondary operator (p) meant for 'libraries' to which it is related. The terms, except proper name (i.e. West Bengal), are written in lower case initials. Index entries will be generated in upper case initials by the computer. Tick mark (") is to be provided for the terms which should appear as Lead (access points) in the index entries.

3) **Generation of Index Entries:** The first index entry will be generated by the computer by pushing the first term of the input string into the Lead position and keeping the remaining terms in the Display position. As soon as any term goes to the Lead position, it is printed in bold type face.

West Bengal

Universities. Libraries. Performance. Measurement

The second index entries will be generated by pushing the second term of the input string into the Lead position and thereby replacing the existing Lead term into the Qualifier position, such as:

Universities. West Bengal

Libraries. Performance. Measurement

Similarly other index entries will be generated as

Libraries. Universities. West Bengal

Performance. Measurement

Performance. Libraries. Universities. West Bengal

Measurement

Measurement. Performance. Libraries. Universities. West Bengal

Note: It can now be seen in the above examples that Lead and Qualifier are separated by a full stop and 2-letter space. The standard separator between two terms in the entry is full stop and one space. The terms in the Display position are written leaving 2-letter space from the left. For over-run of Display in the next line, 4-letter space and for over-run of heading in the next line 8-letter spaces are to be left from the margin.

4) **Generation of supporting reference entries**: 'see' and 'see also' references are generated from semantically related terms taken from a machine-held thesaurus.

5) **Alphabetisation**: All the entries, generated by the process, as stated above, are arranged alphabetically by headings. Under the common heading, displays are organised alphabetically.

Formats of PRECIS Index

Index entries in PRECIS are basically generated in three formats: standard format, inverted format and predicate transformation.

a) **Standard Format**: Index entries in the standard format are generated when any of the primary operators (0), (1), and (2) or its dependent elements appear in the Lead. The process of generation of index entries in the standard format has already been demonstrated under the Section 11.3.4.7 of this Unit.

b) **Inverted Format**: Index entries in the inverted format are generated whenever a term coded by an operator in the range from (4) to (6) or its dependent elements appear in the Lead. The rule relating to the generation of index entries with this format is that—when any of the terms coded either (4), or (5) or (6) or any of their dependent element operators appear in the Lead, the whole input string is read from top to bottom and is written in the Display. However, if the term appearing in the Lead is last term of the input string, then it will be dropped from the Display.

Example: A report on the feminist viewpoint on marriage

Input String:

(2) ✓marriage

(4) ✓feminist viewpoint

(6) ✓reports

Index Entries:

Marriage

– Feminist viewpoint – Reports

Feminist viewpoint

– Marriage – Feminist viewpoint – Reports

Reports

– Marriage – Feminist viewpoint

c) **Predicate Transformation**: When an entry is generated under a term coded (3) that immediately follows a term coded either by (2) or (s) or(t)—each of which introduces an action of one kind or another—the predicate transformation takes place. An input string of this kind is shown below:

Example: Planning of libraries by architect

Input String:

(1) ✓libraries

(2) ✓planning $v by $w of

(3) ✓architects

In order to bring expressiveness in the resulting index entries, the connective codes $v and $w (see 'schema of codes') are attached to the action concept and it results in a compound phrase. The rule relating to the generation of index entries with this format is: when the term coded (3) goes to the Lead, the computer checks the operator assigned to the next preceding term. If that operator is (2) or (s) or (t), the term coded with any of these operators and the term accompanied by the Code $w for upward reading connective (if any) are printed in the Display position instead of Qualifier position. Accordingly, the index entries for the aforesaid input string will be:

Libraries

Planning by architects

Planning. Libraries

By architects

Architects

Planning of libraries

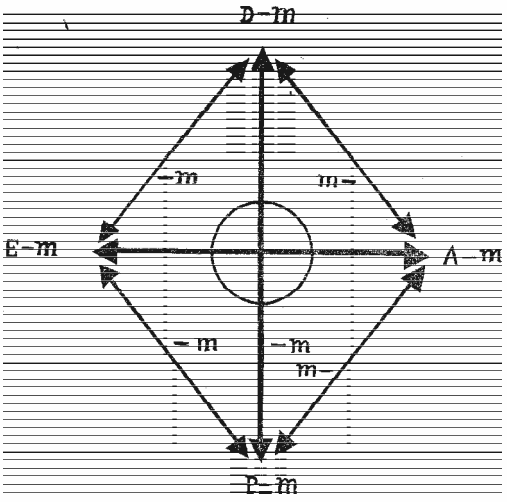
11.3.5 POPSI

All pre-coordinate indexing models are entirely based on the method of facet analysis. Ranganathan pointed out in a paper entitled *'Subject heading and facet analysis' (Journal of Documentation, 20 (3), 1964, p.109-119)* that facet analysis does not depend entirely on notational scheme of classification. The rules of chain procedure, he said, can be so framed as to implement any kind of decision about the sought first heading and the other successive headings in conformity with the principle of local variation. Since then, continuous research on this new line of thinking was going on at Documentation Research and Training Centre (DRTC), Bangalore and a number of papers on Postulate-based Permuted Subject Indexing (POPSI) based on Ranganathan's General Theory of Library Classification came out. Dr. Ganesh Bhattacharyya first explained the fundamentals of subject indexing languages with an extensive theoretical background which ultimately led to the development of newer version of POPSI forming the part of his General Theory of Subject Indexing Languages (GT-SIL). Bhattacharyya developed the POPSI through logical interpretation of the deep structure of subject indexing language (SIL). POPSI drew attention to the helpfulness of adopting a suitable device for ensuring an optimally effective organising classification through the alphabetisation of verbal subject – propositions. It prescribes the use of apparatus words – such as prepositions, conjunctions, participles etc., as and when necessary to communicate the exact meaning of subject propositions. These words are put in parenthesis and they are ignored in alphabetisation. Since the POPSI Index of all verbal entries, filing them in one alphabetical sequence in a unipartite index is made easy.

Major Working Concepts of POPSI

1) **Deep Structure of Subject Indexing Languages (DS-SIL)**

DS-SIL is the logical abstraction of the surface structures of outstanding SILs like Cutter, Dewey, Kaiser and Ranganathan. According to the General Theory of SIL, the structure of a specific SIL has been assumed to be a surface structure of the deep structure of SIL. The DS-SILs has been presented diagrammatically as follows:



It appears from the above diagram that any specific subject may belong to any one of the following elementary categories (D, E, A, P) and modifier:

2) **Elementary Categories and Modifiers**

a) **Discipline (=D)** refers to an elementary category that includes the conventional field of study, or any aggregate of such fields, or artificially created fields analogous to those mentioned above; e.g. Physics, Biotechnology, Ocean science, Library and Information Science, etc.

b) **Entity (=E)** refers to an elementary category that includes manifestations having perceptual correlates, or only conceptual existence, as contrasted with their properties, and actions performed by them or on them; e.g. Energy, Light, Plants, Animals, Place, Time, Environment, etc.

c) **Action (=A)** refers to an elementary category that includes manifestations denoting the concept of 'doing'. An action may manifest itself as Self Action or External Action. For examples: Function, Migration, etc. are Self Actions; and Treatment, Selection, organisation, and Evaluation, etc. are External Actions.

d) **Property (=P)** refers to an elementary category that includes manifestations denoting the concept of 'attribute'—qualitative or quantitative; e.g. Property, Effect, Power, Capability, Efficiency, Utility, Form, etc.

e) **M=Modifier** refers to a qualifier used to modify any one the elementary categories D, E, A and P. It decreases the extension and increases the intension of the qualified manifestation without disturbing its conceptual wholeness. A modifier can modify any one of the elementary categories, as well as two or more elementary categories. Modifiers are of two types:

- **Common Modifiers:** They refer to Space (e.g. Libraries in India), Time (e.g. Libraries in India 19th Century), Environment (e.g. Desert Birds), and Form (e.g. Encyclopedia of Physics). Common modifiers have the property of modifying a combination of two or more elementary categories.

- **Special Modifiers:** A special modifier is used to modify only one of the elementary categories. It may be of Discipline-based, or Entity-based, or Property-based, or Action-based. Special modifiers can be grouped into two types:

  i) those that require a phrase or auxiliary words to be inserted between the term and thus forming a complex phrase, e.g. Cataloguing using computers; and

  ii) those that do not require auxiliary words or phrase to be inserted in between the terms, but automatically form an acceptable compound term denoting Species/Type, e.g. 'Chemical' in 'Chemical Treatment'.

3) **Organising Classification and Associative Classification**

According to the General Theory of SIL, classification is a combination of both organising classification and associative classification. In other words, an indexing system is a combination of both organising classification and associative classification. The tasks

involved in creating an organising classification are the categorisation of concepts and their organising in hierarchies. In organising classification compound subjects are based on genus-species, whole-part, and other inter-facet relationships. Here, classification is used to distinguish and rank each subject from all other subjects with reference to its Coordinate—Superordinate—Subordinate—Collateral (COSSCO) relationships. The result of organising classification is always a hierarchy. In associative classification, a subject is distinguished from other subjects based on the reference of how it is associated with other subjects without reference to its COSSCO relationships. The result of associative classification is always a relative index.

4) **Base and Core**

In the context of constructing compound subject heading, when the purpose is to bring together all or major portion of information relating to a particular manifestation or manifestations of a particular elementary category, the manifestation/category is Base. In case of a complex subject, any one of the subjects can be decided to be the Base subject depending upon the purpose in hand. For example: for a document on 'Eye cancer', 'Eye' is the Base subject in an Eye Hospital Library, and 'Cancer' is to be considered as the Base subject for a Cancer Research Centre.

When the purpose is to bring together within a recognised Base, all or major portion of information pertaining to one or more elementary categories, the category or categories concerned is the Core of the concerned Base. Core lies within the Base, and which one will be the Base or Core depends on the collection or purpose of the library. For example: In DDC, 'Medicine' is the Base, and the 'Human body' and its 'Organs' constitute the Core of the Base.

Features of POPSI'

From the operational point of view, the salient features of POPSI may be grouped under three components: Analysis, Synthesis, and Permutation.

The work of 'Analysis' and 'Synthesis' is primarily based on the postulates associated with the deep structure of SILs for generating organising classification. The task of analysis and synthesis is largely guided by the following POPSI-table. The work of 'Permutation' is based on cyclic permutation of each term-of-approach, either individually or in association with other terms for generating associative classification effect in alphabetical arrangement.

Rules of Syntax

The basic rules of syntax associated with POPSI are:

a) Discipline is followed by Entity, both modified and unmodified.

b) Property follows immediately the manifestation in relation to which it is a Property.

c) Action follows immediately the manifestation in relation to which it is an Action.

d) A Property can have its own Property.

e) An Action can have its own Action.

f) A Species/Type follows immediately the manifestation in relation to which it is a Species/Type.

g) A Part follows immediately the manifestation in relation to which it is a Part.

h) A Modifier follows immediately the manifestation in relation to which it is a Modifier.

The following POPSI Table, like Role Operators in PRECIS, is used in sequencing the component terms for formulating a subject heading

<table>
<tr><td colspan="3">0 Form modifier</td></tr>
<tr><td colspan="3">1 General Treatment</td></tr>
<tr><td colspan="3">2 Phase relation
2.1 General
2.2 Bias
2.3 Comparison
2.4 Similarity
2.5 Difference
2.6 Application
2.7 Influence</td></tr>
<tr><td colspan="2">Common modifiers
3 Time modifier
4 Environment modifier
5 Place modifier</td><td></td></tr>
<tr><td>6 Entity (E)
7 Discipline (D)</td><td>.1 Action (A)
.2 Property (P)
Note: Notations .1 and .2 are preceded by the notation of the manifestation in relation to which it is (A) and (P).</td><td>, Part
. Speciator/Type
—Special modifier
Note: A Species/Type/ Special modifier follows immediately the manifestation in relation to which it is a Species/Type.</td></tr>
<tr><td colspan="3">8 Core I
9 Base (B)
Note: Features relating to Core I and Base (B) are analogous to 6 Entity / 7 Discipline /.1 Action / .2 Property.</td></tr>
</table>

Semantic Relationship in POPSI

Semantic relationship in POPSI is controlled by the vocabulary control device called 'Classaurus'. A classaurus is an elementary category-based (faceted) systematic scheme of hierarchical classification in verbal plane incorporating all the necessary features of a conventional information retrieval thesaurus. Like any faceted classification scheme, a classurus consists of separate schedule for each of the Elementary Categories, namely, Entity, Property and Action, with their Species/Parts and Special Modifiers for each.In each schedule, it displays hierarchical relationship among terms (broader, narrower,

and collateral). Classaurus also contains separate schedules for Common Modifiers like Form, Time, Place, and Environment. Like any thesaurus, each of terms in hierarchic schedule is enriched by synonyms, quasi-synonyms, etc. Unlike a thesaurus, a classaurus does not include associatively related terms (RTs) because of its category-based (faceted) structure. The implication of the faceted structure is that a term in one Elementary Category has a high chance of being non-hierarchically related with another term in another category. The task of showing what is non-hierarchically related to what, and how they are related is left to the care of indexing procedure based on the information contained in the document itself. All the component terms associated with the content of the document are brought together under each approach-term by the permutation technique of POPSI. It is assumed that RTs should not be dictated beforehand by the designer of the classaurus, rather it should be dictated by the content of the document itself. Any term may be related to any other terms depending on the content of the document and hence, RTs should not be determined beforehand.

Steps in POPSI

The main steps in applying POPSI are as follows:

1) **Content Analysis**: Involves identification of different component ideas associated with the content of the document with reference to their elementary categories and modifiers.

2) **Formalisation**: Involves preparing the formalized expression of subject statement obtained on the basis of the results of the step 1 (Content Analysis) according to the rules of syntax.

3) **Standardisation:** Deciding the standard term in the formalized expression of subject statement especially for the term(s) having synonym(s), if any. This step calls for the use of Classaurus.

4) **Modulation:** Involves augmenting the standardized subject proposition by interpolating and extrapolating, as the case may be, the successive superordinates of each manifestation by using standard terms with indication of their synonyms, if any. [*Note: A* Classaurus is the tool to guide the operation in steps 3 and 4 with assurance of consistency in practice].

5) **Entry for Organising Classification**: Involves the preparation of entries for organising classification by inserting appropriate notations for Elementary Categories, subdivisions and modifiers from POPSI Table. Modulated subject proposition with appropriate notations from POPSI Table sorted alpha-numerically will produce organising classification effect by juxtaposition of entries in alphabetical sequence.

   Approach-term Selection: Consists of deciding the approach-terms for generating associative classification effect and of controlling synonyms. The selection of approach-terms may vary for one library to another library depending upon the requirement of users.

6) **Preparation of Entries of Associative Classification**: Involves the preparation of the entries under each approach-term by cyclic permutation of sought terms for generating associative classification effect in alphabetical arrangement.

7) **Alphabetisation:** Involves in arranging all the entries in alphabetical sequence.

Demonstration of the Procedure of POPSI-Basic

0) **Title of the document**:

Use of computers for indexing of educational films in university libraries in India

1) **Content Analysis:**

Library and Information Science = Discipline (D) [Implicit]

University libraries = Entity (E) [Explicit]

Educational films = Part of E [Explicit]

Indexing = A of Part of E [Explicit]

Use = Application phase relation (PR) [Explicit]

Computers = E-based Special modifier (Sm) [Explicit]

India = Common modifier (Cm) of place [Explicit]

2) **Formalisation:**

Library and Information Science (D), University libraries (E), Educational films (Part of E), Indexing (A of Part of E), Use (Application PR), Computers (E-based Sm), India (Cm)

3) **Standardization:**

It is assumed that all the terms in the formalized expression of the subject as shown above are standard terms.

4) **Modulation:**

Library and Information Science (D), Libraries. Academic libraries. University libraries (E), Information sources. Films. Educational films (Part of E), Indexing (A of Part of E), Use (Application PR), Computers (E-based Sm), Asia, India (Cm)

5) **Entry for Organising Classification:**

Library and Information Science 6 Libraries. Academic libraries. University libraries, Information sources. Films. Educational films 6.1 Indexing 2.6 (using) Computers 5 (in) Asia, India

6) **Approach-term Selection:**

The following terms are selected as the approach-terms:

Libraries

Academic libraries

University libraries

Information sources

Films

Educational films

Indexing

Computers

India

7) **Preparation of Entries of Associative Classification:**

LIBRARIES

Library and Information Science 6 Libraries. Academic libraries. University libraries, Information sources. Films. Educational films 6.1 Indexing 2.6 (using) Computers 5 (in) Asia, India

The above organising classification will have to be repeated under each of the following approach terms:

Academic libraries

University libraries

Information sources

Films

Educational films

Indexing

Computers

India

8) **Alphabetisation:**

All the entries are arranged according to the alphabetical order, word-by-word.

POPSI-Specific

The steps in POPSI with illustrative examples as demonstrated above fall within the purview of POPSI-Basic. According to Bhattacharyya, there is no single absolute version of organising or associative classification. POPSI tries to find out what is logically basic, and amenable to systematic manipulation to meet specific requirement. The POPSI-Basic is a product of the application of the GT-SIL and it is readily amenable to the systematic manipulation to generate purpose-oriented specific versions known as POPSI-Specific. POPSI-Specific is always a derivation from the POPSI-Basic according to special decisions and rules to meet specific requirements at the local level. It may be noted here that this approach is totally different from that of earlier contributors of different SILs.

If the purpose is to bring together all or a major portion of information pertaining to a specific topic in a discipline manifesting any of the Elementary categories, the above version of POPSI-Basic can be systematically manipulated to generate the required version of POPSI-Specific. This involves the decision about the 'Base' and 'Core'.

For example, our purpose is to bring together all information pertaining 'Educational films' in one place and hence, 'Educational films' is to be considered as the Base. 'University libraries' and 'Indexing' are to be considered as the Modifier and Action to Base respectively. In view of this, we can prepare the following organising classification entry:

Information sources. Films. Educational films — Libraries. Academic libraries. University libraries 9.1 Indexing 2.6 (using) Computers 5 (in) Asia, India.

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.

5) Discuss the Principle of Context Dependency.

.....

.....

.....

.....

6) Discuss the entry structure and entry format as followed in PRECIS?

.....

.....

.....

.....

7) How do you categorise different operational stages of POPSI?

.....

.....

.....

.....

8) What are the major steps in formulating index entries according to POPSI?

.....

.....

.....

.....

11.4 POST-COORDINATE INDEXING

All indexing systems follow the process of concept coordination to describe the contents of the documents more precisely. We have seen in pre-coordinate indexing, component concepts are coordinated according to the order of significance or citation order by following the syntactical rules of the given indexing language. But the rigidity of the citation order appears to be unsatisfactory to meet the varieties of approaches of all the users. The provision for multiple entries in pre-coordinate indexing by rotating or cycling of the component terms covers only a fraction of the possible number of the total permutations and for this, a large portion of probable approach points is left uncovered. Consequently, the searcher has no choice but to follow the rigid citation order specified by the given indexing language. The above noted problems stemming from the pre-coordination of terms with the rigidity of citation order triggered the development of alternative indexing techniques where the component ideas of a subject are kept separately, uncoordinated by the indexer. Here, concepts/terms are coordinated at the time of searching (i.e. at output stage) by the user. A greater degree of search manipulation

is available in post-coordinate indexing system since the search terms can be coordinated almost in any combination or order to retrieve records of information about the documents as required by the users. The indexing systems which are based on this basic principle are called *Post-coordinate Indexing* or simply *Coordinate Indexing Systems*. Based on this basic principle a number of post-coordinate indexing systems like Uniterm, Optical Coincidence Card, etc. were developed. Among the different types of post-coordinate indexing systems, Uniterm system developed by Mortimer Taube is considered as the most popular post-coordinate indexing model.

11.4.1 Pre-Coordinate Indexing versus Post-Coordinate Indexing

Subjects of documents are not simple. There are compound and complex subjects dealing with multiple numbers of concepts. When there is more than one concept in the document the order in which we cite the concepts and their relationship to one another become important. Both pre- and post-coordinate indexing systems are, by nature, coordinate indexing, but the coordination is done in two different stages. The following table furnishes the points of differences between the two systems:

Pre-coordinate indexing system	Post-coordinate indexing system
Coordination of component terms is carried out at the time of indexing (i.e. at input stage) in anticipation of the users' approach.	Component concepts (denoted by the terms) of a subject are kept separately uncoordinated by the indexer, and the user does the coordination of concepts at the time of searching (i.e. at the output stage).
The most important aspect of this indexing system is to determine the order of significance by following the syntactical rules of the given indexing language.	Rigidity of the significance order is very much absent in this system.
It is non-manipulative. The searcher has no choice but to try to predict the citation order specified by the indexer.	It is manipulative. The searcher has wide options for free manipulation of the classes at the time of searching in order to achieve whatever logical operations are required.
In this indexing system, both the indexer and the searcher are required to understand the mechanism of the system—the indexer for arriving at the most preferred citation order and the searcher for formulating an appropriate search strategy—in order to achieve the highest possible degree of matching of concepts.	This indexing system does not require the indexer and searcher to understand the mechanism of the system. However, the operational aspects n need to be understood by them.

11.4.2 Term Entry System and Item Entry System

In *Term Entry System*, we prepare entries for a document under each of the appropriate subject headings, and file these entries alphabetically. Here, terms are posted on the item (i.e. Term on Item System). In this type of post-coordinate indexing, the number of entries for a document is dependent on the number of terms associated with the thought content of the document. Searching of two files (Term Profile and Document Profile) is required in this system. Uniterm and Peek-a-boo are examples of these.

It is possible to take the opposite approach and make a single entry for each item, using a physical form which permits access to the entry from all appropriate headings. A system which works in this way is called an *Item Entry System*. Here, items are posted on the term (i.e. Item on Term System). In this type of post-coordinate indexing, single entry is made for each item. Item entry system involves the searching of one file (i.e. Term Profile) only. Edge-notched Card is an example of item entry system.

11.4.3 Uniterm Indexing

Uniterm indexing system was devised by Mortimer Taube in 1953 to organise a collection of documents at the Armed Services Technical Information Agency (ASTIA) of Atomic Energy Commission, Washington. Uniterm is a post-coordinate indexing system based on term entry principle. Here, component term (uniterm) is independent of all other terms and serves as a unique autonomous access point to all relevant items in the collection.

Indexing Process

The processes involved in Uniterm indexing are as follows:

1) **Preparation of Document Profile**: Here indexer is required to prepare a card for each incoming document to:
   - assign an accession number or 'address' to each incoming document and record the same on the card,
   - identify the different component terms (uniterms) associated with the thought content of the document,
   - select the uniterms and record them on the card,
   - prepare an abstract for each document and record the abstract on the card,
   - write the bibliographical details of the document on the card, and
   - arrange all the cards as prepared for all incoming documents according to the ascending order of the accession numbers of the documents.

Cards so prepared and arranged according to the ascending order of the accession numbers of the documents are called *Documents Profiles*.

2) **Preparation of Term Profile:** Here indexer is required to
   - prepare a card (term card) for each uniterm,
   - divide each term card into 10 equal vertical columns (from 0 to 9),
   - record the accession number of the document relevant to the uniterm according to the system of terminal digit posting. Terminal digit of the accession number determines the column of its posting, and
   - arrange all term cards according to the alphabetical order of the uniterms.

Term cards so prepared as stated above and arranged according to the alphabetical order of the uniterms are called the *Term Profile* of the system.

Searching

The process of searching in Uniterm system involves the following operations:

- Searcher identifies the different component terms/uniterms associated with the content of her/his queries.
- Searcher, after identifying the component terms/uniterms, pull out the pertinent term cards from the alphabetical deck of the *Term Profile*.
- Term cards thus pulled out are matched to find the common accession number(s). The number(s) common in all such Uniterm cards represent the sum total of the component concept of the specific subject.
- With the help of the common accession number(s), relevant card(s) are pulled out from the *Document Profile* where full bibliographical information of the required document(s) is available.

The main advantage of the Uniterm indexing system is its simplicity and the ease with which persons without much knowledge of subject indexing can handle it. The criticisms against Uniterm system centre around: (a) search time: involves much searching time because of the searching of two files—Term Profile and Document profile; and (b) false drops: possibility of retrieving irrelevant documents due to false coordination of uniterms. For example, searching with the uniterms 'Teachers', 'Students' and 'Evaluation' may retrieve documents on both the subjects, 'Evaluation of students by teachers and 'Evaluation of teachers by students', one of which might be irrelevant to a particular user.

To overcome the problem of false drops, the following post-coordinate searching devices have been used:

1) **Use of Pre-coordinated Terms**

   It is the introduction of pre-coordination to some extent in post-coordinate system in which two or more terms in a subject are bound in place of isolated single term/ uniterm to get rid of false coordination.

2) **Links**: Links are special symbols used to group all the related concepts in a document separately, so that inappropriate combinations of terms are not retrieved. Suppose we have a document (accession number: 243) dealing with two different topics—*Classification of non-book materials and indexing of films.* In order to avoid false coordination like *Classification of films and indexing of non-book materials*, alphabetical symbols, which serve as interlocking device, are attached to accession number to indicate different groups:

Classification	243A
Non-book Materials	243A
Indexing	243B
Films	243B

3) **Roles:** Roles are the indicator digits/symbols attached to the terms at the time of indexing to indicate the role or status or use of the term in a particular context. Here, the possible roles of different terms are identified beforehand and terms are

tagged with these role indicators at the time of indexing. For example, roles developed by the Engineering Joint Council, known as EJC role operators may be attached to *'Television'* to distinguish its functions as the product and tool:

Role	Document
2 [Product / Output]	Manufacturing of *television*
3 [Agent / Tool]	Use of television in education

4) **Weighting:** It is the device of allocating quantitative values to the index terms according to their degree of relevance in the document. Different ways for indicating weights have been suggested. A simple system uses numbers 1 to 3, where 3 indicates maximum weight (i.e. the index term is highly specific and covers an entire major subject of the document), 2 and 1 indicate weights of index terms in decreasing order of their values or relevance in the document.

11.5 KEYWORD INDEXING

Keyword indexing is based on the usage of natural language terminology for generating the index entries. The term 'Keyword' refers to a significant or memorable word (also called 'catchword') that serves as a key in denoting the subject taken mainly from the title of the document (so, it is called *title index)* and sometimes from the abstract or text of the document. Common words like articles (a, an, the) and conjunctions (and, or, but) are not treated as keywords because it is inefficient to do so. This system is also known as N*atural or Free Indexing language.* It is to be pointed out here that the concept of keyword indexing is not new and it existed in the nineteenth century as a 'catchword indexing'. With the introduction of computers in information retrieval in the 1950s, Hans Peter Luhn, an IBM engineer, presented a computer-produced index in 1958 that became known as KWIC (Key Word In Context) indexing.

11.5.1 Key Word in Context Indexing (KWIC)

The production of a KWIC index by H. P. Luhn is the earliest example of an automatic index produced using computers to perform repetitive tasks associated with subject indexing. It was a great step forward in the technique of automatic indexing. Utilizing the capabilities of computers, the KWIC method speedily and with a minimum of intellectual effort produces indexes derived solely from the titles of the documents to be analysed. All significant or key words of titles/title like phrases are alphabetised mechanically, and then printed out in turn following a format which emphasises the selected word. The computer uses the 'stop-word' list in order to ignore all syntactical words such as articles; prepositions etc., and select the remaining words in the title as indexing words. The remaining words of the title are arranged to stand in the context of their original appearance. The use of 'stop-word list' reduces the volume of the index.

The result of the machine manipulation is an index of key terms printed in alphabetical order, together with the text immediately surrounding each term. Each significant word as entry point appears in the margin or a designated middle position while the rest of the title printed on either side. The alphabetical filing is done on the basis of the key word printed in bold letters. *Chemical Titles* (of Chemical Abstract Service) and *BASIC* (Biological Abstracts Subjects in Context) are faithful adaptation of Luhn's KWIC indexing.

Let us consider the following title 'Chemical treatment of cancer in the hospitals of Chennai' to demonstrate the index entries generated according to KWIC indexing:

Cancer in the hospitals of Chennai / Chemical treatment of	614
Chemical treatment of cancer in the hospitals of Chennai	614
Chennai / Chemical treatment of cancer in the hospitals of	614
Hospitals of Chennai / Chemical treatment of cancer in the	614

Annotation

a) Title in the above KWIC index has been rotated in such a way that each keyword serves as the approach term and comes in the beginning by rotation followed by rest of the title;

b) Last word and first word of the tile are separated by using a symbol say, stroke [/] (sometime an asterisk '*' is used) in an entry. In some computer-produced KWIC indexes, keywords are positioned at middle of the entry;

c) Keywords are printed in bold type face to bring prominency in the approach term;

d) Identification / location code 614 is given at the right end of each entry; and

e) Entries are arranged alphabetically by keywords.

Advantages and Disadvantages of KWIC

KWIC method offers the following advantages: (1) the detection of formulaic expressions and repeated word patterns; (2) simplicity; (3) speed; (4) maximisation of computer use; and (5) minimisation of the indexer's role.

The most common type of complaint against the KWIC indexing method is the lack of terminology control as it is entirely dependent upon titles/abstract/text of the document. Apart from this, KWIC has basically two problems: (i) KWIC shows sentences which contain distant dependency as different context, and (ii) KWIC also shows sentences which have different word order as different context. The effects of computer's inability to resolve these problems led to the redundancy, scatter of references throughout the index, haphazard groupings and retrieval losses because the user is forced to guess at the terminology the author actually used. The disadvantages of KWIC-type index can be summarized as follows:

— Large number of index entries under a given keyword, which provokes difficulties in searching;

— Lack of significant words in titles (therefore the title and the abstract are often used as a source for indexing to increase the depth and range of indexing);

— No cross references, which make it difficult to find synonyms, spelling variants and inflections in the index;

— Relatively high computing time, due to superfluous non-significant index entries;

— No combination of keywords;

— Lack of consistency in the indexing terms, because different authors can use different form of words to communicate the same idea, or give different meanings to the same word or phrase;

— References are not grouped under a convenient heading;

— Redundancy of the index.

11.5.2 Variations of Keyword Indexing

A number of varieties of keyword index appear in the literature and they differ only in terms of their formats but indexing techniques and principle remain more or less same. Some important versions of keyword indexing are discussed below:

Key Word Out of Context (KWOC)

In KWOC, each index word is extracted from its context and printed separately in the left hand margin with the unmodified title in its normal order printed to the right. For example:

Cancer Chemical treatment of cancer in the hospitals of Chennai 614

Chemical Chemical treatment of cancer in the hospitals of Chennai 614

Chennai Chemical treatment of cancer in the hospitals of Chennai 614

Hospitals Chemical treatment of cancer in the hospitals of Chennai 614

Sometimes, keywords are positioned and printed as heading and the title is printed in the next line under the heading instead of the same line as shown above.

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

It has been observed that the dependency of keyword indexing on titles sometimes fails to represent the thought content of the document co-extensively. In order to solve this problem, KWAC index came into being. In KWAC, the keywords of the title are enriched 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 generate further index entries. KWAC is nothing but the enrichment of KWIC or KWOC. Further enriched KWIC or KWOC gives index entries wherein additional terms are inserted into the title or added at the end. This involves intellectual effort in the selection of additional terms. CBAC (Chemical Biological Activities) of BIOSIS uses KWAC index where title is enriched by another title like phrase formulated by the indexer.

11.5.3 Double KWIC

It is another improved version of KWIC. Double KWIC index is constructed in the following way:

a) The first significant word in a title is extracted as a main index term and replaced by an asterisk (*) to indicate its position in the title.

b) The remaining words in the title are then rotated, so as to permit each significant word to appear as the first word of a wrap-around subordinate entry under the main index term. Steps 1 and 2 are repeated until all of the titles of a given bibliographic listing are processed. The index entries so created are then sorted alphabetically, both with regard to main terms (primary sort) and subordinate terms (secondary sort). Main index terms are not restricted to single words, but may consist of multi-word terms derived from contiguous sets of words in the titles, for example:

ANNUAL REPORT
EDITORIAL:*
1966=..... F 1
ANNUAL REVIEW
BOOK REVIEW:* OF INFORMATION SCIENCE AND TECHNOLOGY= B3-2.
INFORMATION SCIENCE AND TECHNOLOGY= BOOK REVIEW:*OF B3-2
SCIENCE AND TECHNOLOGY= BOOK REVIEW:* OF INFORMATION B3-2
TECHNOLOGY= BOOK REVIEW:* OF INFORMATION SCIENCE AND B3-2

11.5.4 Other Versions

In addition to the above variations in keyword indexing, a number of varieties of keyword index are available and they differ only in terms of their formats but indexing techniques and principles are more or less the same. They are:

i) **KWWC (Key-Word-With-Context) Index:** In KWWC, 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:** The KEYTALPHA is just modified form with key terms arranged alphabetically. 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 along with the key words, the names of authors are also treated as keywords and thus indexed accordingly. Thus, it appears that WADEX satisfies both the author and subject. WADEX is used in 'Applied Mechanics Review'. AKWIC (Author and keyword in context) index is another version of WADEX.

iv) **KLIC (Key-Letter-In-Context) Index:** This type of index only takes fragmented word (i.e. key letters), instead of the full word, either at the beginning or at the end of the entry. In this system, the key letters forming the part of the word are specified and the computer retrieves any term containing that letters either at the beginning or at the end of the word. KLIC indexes are almost unknown today, the Chemical Society (London) published a KLIC index as a guide to truncation.

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.

9) What do 'Pre-' and 'Post-' signify in Pre-coordinate and Post-coordinate indexing systems?

.....

.....

.....

.....

10) What are the search devices used to avoid false coordination of terms in Post coordinate indexing?

.....

.....

.....

.....

11) Mention the different varieties of keyword indexing.

.....

.....

.....

.....

11.6 COMPUTERISED INDEXING

You already came to know that information retrieval deals with the problems related with the storage, access and searching of information sources by persons in need of information. In this digital network era, information sources are growing at an exorbitant rate, available in many forms and formats, and accessible through various channels. Moreover, recent advancements in Information and Communication Technology (ICT) help in integration of different information sources and process them on a larger scale. The results of ICT applications in library activities are related with the development of wider and efficient information services. Remote database search service (both bibliographic and full-text databases) is possibly one of the most prominent products of ICT-enabled library services.

But you will be surprised to know that investigations of computerised indexing date back to the late 1950s. The earliest and most primitive form of computerised indexing method relying on the power of computers was invented by Hans peter Luhn, an IBM engineer, who in 1958 produced what became known as KWIC (Key Word In Context) indexing. Luhn reported his system in 1960 which has already been discussed under the sub-section 11.5.1. The system was based just on simple, mechanical manipulations of terms derived from document titles. Related forms are the Permuterm Subject Index and the KeyWord Plus known from ISI's citation indexes (this last system is based on assigning terms from cited titles). Computers are now increasingly used in aiding indexing, relying on stored dictionaries of synonyms and homonyms, lists of chemical compounds, plants and animals, etc. They are also used for automatically arranging entries in alphabetical order or subordinating subheadings and cross references in exact sequence under a heading, and performing many other functions that previously had to be done manually and therefore were quite expensive and often subject to errors. After the introduction of Unicode (a 2 Byte-oriented encoding standard that can represent all characters of all scripts of the world) as text encoding standard (in late nineties), computerised indexing systems are able to store, process and retrieve multilingual documents available in different scripts.

11.6.1 Meaning and Features

Computerised indexing can be defined as the process whereby a computer is used to process a natural language text that is already in machine-readable form so that indexing terms are allocated to its content without direct human intervention.

The features of computerised indexing are:

1) Computerised indexing starts with words.
2) Word association prompts the linking of target words in a search statement.
3) Computers scan text and create 'inverted file' which associates words in the file with position in the texts.
4) Matches words in a search statement against 'inverted files' to identify texts that have words in common.

5) Computer algorithms are used to carry out the above operations.

6) Humans do the programming and set the parameters for indexing.

7) Computation techniques used include word frequency and keyword analysis.

8) Computerised indexing cannot replace human or manual indexing, rather complimentary.

11.6.2 Manual Indexing versus Computerised Indexing

Manual Indexing	Computerised Indexing
• Human indexer analyses texts and selects terms for indexing. • Human indexer interprets and encodes text, and makes inferences and judgment in selecting index terms judiciously. • Semantic, syntactical as well as contextual considerations govern the selection of indexing terms • Disagreement among indexers on the determination of the subject of a document as the process of determination of subjects may variy from one indexer to another indexer. • Selected index terms less in numbers. • It is time consuming. • It is expensive. • It is very difficult to maintain consistency in indexing. • Relevancy of results are ensured • Generally maintains a balance in recall-precision.	• Computer analysis of texts has not achieved the reliability of human analysis. Computer analysis of texts is carried out by following the human instructions in the form of a computer programming. Computers can select words by employing statistical technique. • No text interpretation is possible as computer cannot think and draw inferences like human indexer. It can select or match keywords which are provided as input text. • Computer algorithms are drawn to select, or exclude a term by following the rules of semantic, syntactical and contextual connotations, like human indexer. • Determination of the subject of the document is a mechanical process, based on what terms appear frequently and/or prominently in the text—i.e. more frequently a term occurs in a document, the more likely it is that the document is about that term. • Selected index terms are more in number. • It takes less time. • Index entries can be produced at lower cost. • Consistency in indexing is maintained. • Relevancy of results are not always ensured. • Generally shows high recall and low precision.

11.6.3 Advantages and Disadvantages of Computerised Indexing

Advantages of computerised indexing are as follow:

- It is as effective as human indexing;
- It is cost effective compared to expensive human indexing;
- Maintains consistency in indexing;
- Indexing time is reduced;
- Help searchers find information quickly;
- Can be applied to large volumes of texts where human indexing becomes impossible (e.g. Indexing web pages);
- Retrieval effectiveness can be achieved.

Disadvantages are

- Not flexible;
- Not precise when looking at unique materials;
- Not able to adapt new terminology;
- Not able to do the conceptual analysis of the content of the document;
- Not a term occurs several times in a document will always be a significant term.

11.6.4 Components of Computerised Indexing System

A typical computerised indexing system has four major components:

1) **Database**

Database acts as heart of a typical computerised indexing system. Bibliographic databases which deal with metadata of bibliographic entities (e.g. author, title, subject etc.) include two parts. The first part is **sequential file** (a combination of fields —> records —> database) and **inverted file** (indexes to sequential file). Full-text databases, apart from the above two parts also include field-less information entity (source object in different formats like Web page (HTML), PDF file, Doc file etc).

2) **Search Process**

Database determines what can be retrieved, whereas search mechanism determines how information stored in databases can be retrieved. Efficiency of search process is very important factor for a computerised indexing system. Search process of a typical computerised indexing system provides two sets of retrieval techniques – basic retrieval techniques (Boolean operators – AND, OR, NOT; Relational operators - >, <, =, >=, <=; and Positional search operators – NEAR, ADJ, NEARx etc.) and advance techniques (Weighted searching, Fuzzy searching etc.).

3) **Language of Indexing**

Search mechanism determines what retrieval techniques will be available to searchers, whereas computerised indexing language determines the flexibility in (1) document representation; and (2) query representation. Computerised indexing language may

be grouped as natural language and controlled vocabulary (classification, subject heading and thesauri).

4) **User Interface**

It is a layer of interaction between users and computerised indexing system. Efficiency of user interface depends on mode of interaction, display features, online help, provision of feedback etc. It is considered as the human dimension of computerised indexing system.

11.6.5 Categories of Computerised Indexing Systems

Since the time of H.P. Luhn (i.e. 1950s) different computerised indexing systems have been evolved to address information demands of different user groups. These indexing systems may be categorised as follows –

Category I: Online Database Indexing

These computerised indexing systems allow users to search databases located in remote places. Here computer technologies are applied to process, store, retrieve records and communication technologies help in accessing records from centralised databases. Generally, these IR systems include bibliographical, numeric, full-text and multimedia information bearing objects. The online IR systems played a major role in the development of computerised indexing systems over the years.

Advantages: Sophisticated retrieval techniques, cross-database searching, use of integrated vocabulary control devices and many more.

Limitations: Heavy initial investments, need of intermediary in searching etc.

Category II: Optical Disk based Database Indexing

These computerised indexing systems emerged from online IR systems. These IR systems offer subset of data (bibliographical, numeric, full-text, and multimedia) through the optical media (like CDROM, DVDROM etc.). These IR systems adopted almost all the retrieval techniques of online IR systems but the end users, rather than the intermediary, does most of the searching in optical disk based IR.

Advantages: Low running cost, end-user friendly, browsing and searching facilities, sharing of databases on LAN etc.

Limitations: Delayed updating (updating frequency ranges from quarterly to bi-annually) and restricted access (remote searching is not possible beyond LAN).

Category III: OPAC based Indexing

An OPAC is outer / external form of library catalogue. OPACs are presently considered as IR systems with their own characteristic features. For example, an OPAC is essentially local (serve library resources of one or more institutes) but can act as global information entity by integrating other IR systems e.g. CDROM databases, e-journals access, Z 39.50 based searching of other library catalogues etc.

Advantages: Supports field-level search (i.e. search by author, title etc.), and subject access points, provides sophisticated retrieval techniques, and facilitates integration of different information resources into a single search interface.

Limitations: Lack of federated search options, absense of multi-lingual user interface etc.

Category IV: Indexing Internet Resources

The first three computerised indexing systems are dealing with structured objects (fields and field values) whereas the last one deals with unstructured information bearing objects (textual objects without metadata). Internet based computerised indexing systems are generally based on automatic harvesting devices such as robots, spiders, crawlers etc. for finding and gathering of information resources available in publicly indexable Internet (see section 11.7).

Advantages: Free indexing services like search engines, subject directories amd meta search engines, quick access to huge information sources, supports end-user searching, simple to retrieve documents.

Limitations: High recall and very low precision, cross-disciplinary semantic drift, relevancy of results not ensured etc.

11.6.6 Comparison of Computerised Indexing Systems

In the previous section we discussed different categories of computerised indexing systems by covering features of the major systems including the relative advantages and disadvantages of those indexing systems. In this section we are going for a comparison of major computerised indexing systems on the basis of a set of defined parameters.

Features	Online Databases	Optical Disk based Databases	OPAC	Internet
Contents	Mainly textual objects; Bibliographic metadata	Bibliographic metadata, Fulltext and Multimdia	Bibliographic metadata	Text, Images and Multimedia
Retrieval approach	Searching	Searching and browsing	Searching and browsing	Searching and browsing
Indexing	Metadata and Keywords in abstract	Metadata and Keywords in full-text	Metadata and Keywords in TOC	Limited metadata, Keywords in full-text
Retrieval techniques	Basic and Advanced (Limited)	Basic and Advanced (Limited)	Basic and Advanced (Full)	Basic and Advanced (full) including Fuzzy
Search modification	Yes	Yes	No	Yes
Search method	Indexing based (expert search)	End-user search	End-user search	End-user search
Output / Display	Traditional ranking	Traditional ranking	Traditional ranking and Limited modern ranking	Different modern ranking facilities
Use of controlled vocabulary	Yes	Yes	Yes	No
Quality control mechanisms	Yes	Yes	Yes	No
Multilingual search	No	No	Yes	Yes

However, the trend of in the domain of IR system is convergence. CDROM based systems are integrated with OPACs, OPACs are linking online databases, document delivery services, and other resource discovery services. MARC 21 bibliographic format includes field 856 for encoding URLs of Internet resources. The Web is becoming the platform for convergence of different IR systems e.g. Web-OPACs are linking open databases (information mashup) and acting as the gateways for local and global information resources to support users.

11.6.7 Index File Organisation

You must have noticed that all four major computerised indexing systems depend on database as core component. Database, on the other hand depends on two basic parameters – method for handling sequential file and method for handling inverted file. Sequential file management ensures organisation of records in a database whereas inverted file management helps in retrieving records from database against queries. Therefore, design and development of the computerised indexing system mainly depends on:

1) Method used for organising records in the file (file organisation), and

2) Method used for searching a record in the file (search techniques).

Further, these two factors are related in a sense that more efficient the file organisation, more efficient the searching.

In computerised indexing, each document is represented by a record that a computer can read and manipulate. Typically, these records contain data identifying an item and a set of index terms to provide subject access. A basic problem in computerised indexing is storing of files consisting of records, each record having an identical format. A record format consists of a list of fields, with each field processing a number of characters and having a fixed data type. A record also consists of values for fields. The complexity of organising a file for storage depends on the operations we intend to perform on the file.

A file can be categorised as a logical file or a physical file. A logical file is the one perceived by an application program, it may be different from the one which is stored in storage units. Logical files are thus abstracts groups of data. Actual data are stored using physical files. A physical file is a set of data stored on a physical medium such as disk, tape, etc. It contains a number of data subsets, called physical records, which have an identical layout.

Some of the important logical file organisations are discussed below:

a) **Sequential File**: The most common organisation of records in a file is the sequential file organisation. It is the document file, which contains document records in their normal form—the form in which they are sequentially entered into the database. Here, document records are stored one after another in the computer memory—this is actually the virtual structure of the database file.

b) **Inverted File**: An inverted file is a computer file in tabular format, in which rows represent documents and columns represent words. Intersections of rows and columns are marked when certain documents contain certain words. At the point of retrieval, the computer scans the entire inverted index for documents which contain the words in the search query. In an inverted file organisation, two files are always maintained: sequential file and inverted file. An inverted file contains all the potential index terms arranged alphabetically, drawn automatically from the

document records according to indexing technique adopted for the purpose. Each index term in the inverted file is associated with the record number(s) in which the index term occurs and it links with the list of records that represents documents. Thus, for each index term in the database the inverted file contains an entry along with a reference list which specifies position(s) in the database where the term appears. Each term may occur in a number of documents.

c) **Indexed Sequential Files**: An indexed sequential file is also an inverted file in which every record in the source file/document file.

d) **Chained Files:** It's a special type of inverted file organisation which supports dynamic and multiple record linking to ensure simple updating process and quick response against query.

e) **Tree Structured Files**: Here records in a file based on the keys can be organized using a tree structure. If the record is too large, only the keys (in the directory) are organized in the tree structure.

f) **B-Tree**: Bayer and McCreight suggested a method of storing keys on a disk by developing what is called a page. Here a block of storage of fixed size used to transfer information between main storage and direct access storage. This is called B-Tree indexing system that allows easy retrieval, insertion, and deletion of records.

11.6.8 Methods of Computerised Indexing

The first step in indexing, both manual and computerised indexing, is to decide on the subject matter of the document. In manual indexing, the indexer would consider the subject matter in terms of answer to a set of questions such as "Does the document deal with a specific product, condition or phenomenon?" Computerised indexing follows a set of processes of analysing frequencies of word patterns and comparing results to other documents in order to assign to subject categories. This requires no understanding of the material being indexed therefore leads to more uniform indexing but this is at the expense of the true meaning being interpreted. A computer program will not understand the meaning of statements and may therefore fail to assign some relevant terms or assign incorrectly. Human indexers focus their attention on certain parts of the document such as the title, abstract, summary and conclusions, as analysing the full text in depth is costly and time consuming. A Computerised system takes away the time limit and allows the entire document to be analysed, but also has the option to be directed to particular parts of the document.

The second stage of indexing involves the translation of the subject analysis into a set of index terms. This can involve extracting from the document or assigning from a controlled vocabulary.

Statistical Method

Statistical method involves taking words directly from the document. It uses natural language and lends itself well to automated techniques where word frequencies are calculated and those with a frequency over a pre-determined threshold are used as index terms. A stop-list containing common words like articles, conjunctions, prepositions and pronouns are excluded as index terms using a 'stop word' file. Automated extraction of terms may lead to loss of meaning of terms by indexing single words as opposed to phrases. Although it is possible to extract commonly occurring phrases, it becomes more difficult if key concepts are inconsistently worded in phrases. The following statistical methods are adopted in measuring the word significance:

a) **Term Frequency method**: In this method, terms (other than common words) to be indexed are those occurring either very frequently (indicating concepts dealt with) in a text or very seldom (indicating a topic mentioned expressly only once or twice in the title or first paragraph but then being referred to by 'it' or 'this' and the like). This method is based on the frequency of occurrence and co-occurrence of terms, using probabilistic model.

b) **Relative Frequency Method:** Terms that occur infrequently may be highly significant for example a new drug may be mentioned infrequently but the novelty of the subject makes any reference significant. One method for allowing rarer terms to be included and common words to be excluded by automated techniques would be a relative frequency approach where frequency of a word in a document is compared to frequency in the database as a whole. Therefore a term that occurs more often in a document than might be expected based on the rest of the database could then be used as an index term, and terms that occur equally frequently throughout will be excluded.

c) **Term weighting method**: Not all the words have the same significance level. Words occurring with high frequency in a document are better discriminators than words of low frequency. Hence, term frequency method (how often a term is use) is sometimes coupled with term weighting in which different degrees of importance is assigned to terms on the basis of what terms are used in a search request or on the basis of where and how terms appear (e.g. in the title, in an abstract, or in the first and/or last paragraph of a text). For this, term weighting schema is prepared. This method is based on statistical principle.

Linguistic Method

A quite a different approach to computerised indexing is by syntactic and semantic analysis of text. The syntactical analysis identifies the grammatical role and relation among the words in the sentence. It is concerned with automatic recognition of significant word order in a phrase or sentence and with inflections, prefixes, and suffixes that indicate grammatical relationships. Semantic analysis approach seeks to analyse noun phrase automatically with the aid of stored dictionaries and other linguistic aids. Syntactic and semantic analyses are often used in conjunction. Natural Language Processing (NLP) is based on these two methods. Linguistic ontologies are used in NLP to assist in the analysis of natural language text.

Artificial Intelligence (AI) based Indexing System

The application of artificial intelligence in information retrieval research basically involves processing of source text to identify the roles of words and phrases and the relationships between them. The results of this processing are used to identify appropriate indexing expressions. The following indexing system researches based on AI are evident:

a) **Natural Language Processing (NLP) based Indexing System:** The basic idea is to process the text of documents to generate indexing terms. Methods focusing on the lexical level, attempting to identify grammatical classes or parts of speech of individual words together with machine readable dictionaries to index documents. At the syntactic level, there have been attempts to examine the interconnections between words. It uses both co-occurrences of pairs of terms and threshold distances (number of words between two terms). The ideas of co-occurrence and threshold distance may be seen as an attempt to address the issue of coordination (relationships between subject descriptors). This approach is based

upon a network of expertise or knowledge, the elements of which are associated with the documents represented in the database.

b) **Expert System based Indexing System**: Expert systems have been used in information retrieval research in an attempt to replace what is usually referred to as the expert intermediary. The expertise referred to in this case is that necessary to construct a search appropriate to the user's needs and the functioning of the system. The components therefore include knowledge of how to interrogate the system, how to identify the appropriate search terms and how to link these terms. An expert system shell consists of an appropriate conceptual hierarchy and a dictionary of the appropriate subject area. This is created by analysis of subject information by experts in the subject area. Associated with each node in the hierarchy is relationship information for navigating the system and document references. Searching basically consists of browsing the network. The network nodes are linked by different kinds of relationship.

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.

12) Why should we use computerised indexing system in libraries?

.....

.....

.....

.....

13) What are the major components and categories of computerised indexing system?

.....

.....

.....

.....

14) Distinguish between Optical Disk based Indexing and Online Database Indexing.

.....

.....

.....

.....

15) What are the different methods used for organising records in the index file?

.....

.....

.....

.....

16) Highlight the different methods used in computerised indexing.

.....

.....

.....

11.7 INDEXING INTERNET RESOURCES

You already know in sub-section 11.6.5 that Internet is evolving as the largest information system in the world and Internet based indexing system is now considered as a major category under computerised indexing system. The global growth in popularity of the World Wide Web (WWW) has been enabled in part by the availability of browser based search tools which in turn have led to an increased demand for indexing techniques and technologies. Information resources accessible through the Web are very much different from the bibliographic records of a conventional system. The Web presents information access problems orders of magnitude greater than any encountered before. The major defect of the Internet as an information source, apart from its sheer size, is the fact that it lacks any form of quality control. Current chaotic situation caused by the "every man his own publisher" phenomenon. Publishers of scholarly books and journals apply reviewing/refereeing procedures that are, at least to some extent, effective in eliminating the most worthless of what is written. The published indexing and abstracting services provide the next level of quality filtering, mostly by choosing the journals, report series, or other publications that they cover on a regular basis. It is obvious that human professional indexing of the entire Web is completely impractical. Selective professional indexing, of course, is possible.

There have been several attempts to organise the resources on the WWW. Some of them have tried to use traditional Library Classification Schemes such as the Library of Congress Classification, the Dewey Decimal Classification and others. However there is a need to assign proper subject headings to them and present them in a logical or hierarchical sequence to cater to the need for browsing.

When we talk about indexing Internet resources, we really mean indexing Web resources or simply Web indexing. It centres round on the following:

a) search engine indexing of the Web,

b) creation of metadata,

c) organisation of Web links by category, and

d) creation of a Website index that looks and functions like a back-of-book index.

11.7.1 Search Engine Indexing

Internet search engines are special sites on the Web that are designed to help people find information stored on other sites. A search engine searches a database of information on the web. It is a tool to help users to locate information available via Internet. There are differences in the ways various search engines work, but they all perform three basic tasks:

- They search the Internet — or select pieces of the Internet — based on important words.
- They keep an index of the words they find, and where they find them.
- They allow users to look for words or combinations of words found in that index.

To find information on the hundreds of millions of Web pages that exist, a search engine employs special software robots, called spiders (other names for these programs are crawler, worm, wanderer, gatherer, and so on), which traverse the web, following links between pages. It builds the list of the words found on Web sites. When a spider is

building its lists, the process is called Web crawling. In order to build and maintain a useful list of words, a search engine's spiders have to look at a lot of pages.

How does any spider start its travels over the Web? The usual starting points are lists of heavily used servers and very popular pages. The spider will begin with a popular site, indexing the words on its pages and following every link found within the site. In this way, the spidering system quickly begins to travel, spreading out across the most widely used portions of the Web.

When the spider looked at an HTML page, it took note of two things:

- The words within the page
- Where the words were found

Words occurring in the title, subtitles, meta tags and other positions of relative importance were noted for special consideration during a subsequent user search. Meta tags allow the owner of a page to specify key words and concepts under which the page will be indexed. The meta tags can guide the search engine in choosing which of the several possible meanings for these words is correct.

Once the spiders have completed the task of finding information on Web pages, the search engine must store the information in a way that makes it useful. There are two key components involved in making the gathered data accessible to users: (1) The information stored with the data, and (2) The method by which the information is indexed.

In the simplest case, a search engine could just store the word and the URL where it was found. To make for more useful results, most search engines store more than just the word and URL.

Search Engine Categories

Search engines may be divided into the following main categories:

- **General Search Engine**: It covers a range of services and compiles their own searchable databases on the web. Examples: Google, Alta Vista, etc.
- **Regional Search Engine**: It refers to country specific search engine for locating varied resources region-wise. Examples: Euro Ferret (Europe), Excite uk (UK), etc.
- **Meta-search Engine**: A meta search engine does not use crawler for compiling their own searchable database. These search engines utilise databases maintained by other individual search engines. When a query is put before this type of search engine, it forwards that query to other search engines. Examples of such search engines are: Dog pile, Ask Jeeves, Inference Find, MetaCrawler, Profusion, Surfwax.
- **Subject Specific Search Engine**: It does not attempt to index the entire Web. Instead, it focuses on searching for Websites or pages within a defined subject area, geographical area or type of resource. Examples: Geo index (Geography / Environmental science), Biochemistry Easy Search Tool (Biochemistry). Because this specific search engine aims for depth of coverage within a single area, rather than breadth of coverage across subjects, they are often able to index documents that are not included even in the largest search engines databases. Some examples of subject specific search engines are: , Bioweb (Biotechnology), Scirus (Sc. & Tech.), Medical world search, Health A to Z (Medical Sc.), Math Search

(Mathematical Sc.), Agri Surf (Agricultural Sc.), Law Crawler (Law), KHOJ (India specific search engine), etc.

- **Directory-based Search Engines**: Due to explosion of information over Internet it is felt that the results fetched by search engines are often mixed with unwanted ones. Subject directories, unlike search engines, are created and maintained by human editors, not electronic spiders or robots. Directory editors typically organise directories hierarchically into browsable subject categories and sub-categories. The resources they list are usually annotated. Directories tend to be smaller than search engine databases, typically indexing only the home page or top level pages of a site. They may include a search engine for searching their own directory. It enables the searcher to move from menu to menu, making one selection after another until he gets to the level where the chosen sites are enlisted. These directories offer access to the information that has been classified into certain categories. Examples of such search engines are: Yahoos, Google, Look smart, Magellan, Open Directory, and Project.

Salient points of differences between Subject Directories and Search Engines are as furnished below:

	Subject Directory	Search Engine
Coverage	It is a categorised list of websites with brief description, based on submission by Web site owners, scrutinised and edited by professional editors. Contents coverage in a subject directory are fewer as compared with a search engine.	A search engine indexes all the information on the entire web automatically. It deals with specific piece of information, not categories. Hence, contents coverage in a search engine is higher than the search engine.
Browse	Allows searchers to browse resources in the home page of a Web site by predefined subject categories. From there, searcher can explore the site for more specific information.	Does not allow searchers to browse resources by predefined subject categories. It takes the searcher to the exact page on which the words and phrases he/she is looking for appear.
Searching	Searches titles and annotations of categorised resources.	Searches full-text of web pages.

11.7.2 Subject / Information Gateway

With the advent Internet many libraries are looking forward to go online with Internet. Often they are finding the information available over Internet is enormous. For the same they have devised subject-based portals which are known as Subject Gateways. Major idea of Subject gateways came with the inefficiency of search engines as they fail to give pin-pointed information. They employ subject experts and information professionals to select, classify and catalogue Internet resources to aid search and retrieval for their users.

A subject/information gateway is a web site that provides searchable and browsable

access to online resources focused around a specific subject. Subject gateway resource descriptions are usually created manually rather than being generated via an automated process. There are two kinds of gateways: library gateways and portals. Library gateways are collections of databases and informational sites, arranged by subject that have been assembled, reviewed and recommended by specialists, usually librarians. These gateway collections support research and reference needs by identifying and pointing to recommended, academically-oriented pages on the Web. Some notable examples of library gateways are Librarians' Index to the Internet, Internet Public Library, Academic Information, WWW Virtual Library, Digital Librarian, Infomine, etc.

11.7.3 Semantic Web

Search engines usually employ statistical methods like frequency of occurrence of words, co-occurrence of words, etc. which retrieve a number of irrelevant hits against a search on the Web. Though some search engines like Google and Yahoo use human edited entries; still they come up with a large number of wrong hits. Tim Berners-Lee introduced the concept of Semantic Web to extend the web with semantic information. The idea behind Semantic Web is to develop such technologies that make the information more meaningful for the machines to process, which in turn makes search and retrieval of information more effective for searchers. The conception of Semantic Web is characterized by developing tools and technologies like languages, standards and protocols so that the Web becomes meaningful.

Most of the technologies involved in the development of the Semantic Web are still in their infancy. Some of them already in use are the URIs (for identifying documents uniquely and globally), XML (for structuring the data semantically), RDF (to base the structure of the documents on a common model base), Ontologies (to define the objects/entities and the interrelations between these objects/entities), etc.

11.7.4 Taxonomies

Taxonomies can be considered another tool useful for organising web-based information. For example, taxonomies provide an excellent means for organising subject-specific information into an easily navigable format. The utility of taxonomies for displaying web information can also be found when examining Yahoo's site – which uses a taxonomy/subject hierarchy to classify its indexed information.

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.

17) What do you mean by indexing Internet resources?

.....

.....

.....

.....

18) What are the different types search engines used in indexing Internet resources?

.....

.....

.....

19) Discuss the methods of search engine indexing.

.....

.....

.....

20) What is Semantic Web?

.....

.....

.....

11.8 SUMMARY

In this Unit we have dealt with different techniques of subject indexing. It begins with a brief discussion on derivative indexing and assignment indexing and is followed by the discussion on different types of pre- and post-coordinate indexing systems. We cannot understand the pre-coordinate indexing properly without being aware of the contributions of C.A. Cutter and J. Kaiser. For this, principles and processes of subject indexing techniques as enunciated by Cutter and Kaisers are discussed. Major pre-coordinate indexing systems like Chain Indexing, PRECIS, and POPSI are discussed with reference to their principles, syntactical and semantic aspects, entry structure and system of references. Objective conditions that led to the development of post-coordinate indexing and its differences with pre-coordinate indexing are explained. Term entry system and item entry system forming parts of the post-coordinate indexing system are also discussed with an emphasis on the operational stages of Uniterm indexing system. Different varieties of keyword indexing are explained. Computerised indexing techniques are explained in terms of its meaning, features, differences with manual indexing, advantages and disadvantages, components, categories, index file organisation and different methods associated with the generation of index entries with the aid of computers. Indexing internet resources with particular reference to search engine indexing and other associated concepts are discussed briefly at the end of this Unit. All indexing techniques are demonstrated with illustrative examples.

11.9 ANSWERS TO SELF CHECK EXERCISES

1) Derived indexing is a method of indexing in which a human indexer or computer extracts from the title and/or text of a document one or more words or phrases to represent subject(s) of the work, for use as headings under which entries are made. It is also known as *extractive indexing*.

   Assigned indexing is also known as 'Concept indexing', because what we are trying to do is to identify concept(s) associated with the content of each document. It is a method of indexing in which a human indexer selects one or more subject headings or descriptors from a list of controlled vocabulary, instead of the title and/or text of a document, to represent the subject(s) of a work.

2) J. Kaiser in his "Systematic Indexing", published in 1911, solved the problems indexing compound subjects through classificatory approach. He pointed out that compound subjects might be analysed by determining the relative significance of the different component terms of compound subject in terms of two fundamental categories: (1) Concrete and (2) Process. According to Kaiser, Concrete refers to Things, place and abstract terms, not signifying any action or process; e.g. gold,

India, Physics, etc. Process refers to (a) Mode of treatment of the subject by the author; e.g. *Cataloguing of films*; (b) An action or process described in the document; e.g. *Cultivation of rice*; and (c) An adjective related to the concrete as component of the subject; e.g. Strength of metal. Kaiser laid a rule that a '*Process'* should follow *'Concrete'*.

3) The concept of 'chain' as the foundation of chain indexing is a structural manifestation of a subject, which refers to the parts constituting a subject and their mutual interrelationship. It is a modulated sequence of sub-classes or isolate ideas.

   Since the chain expressed the modulated sequence more effectively in a notational classification of subjects, this method takes the class number of the document concerned as the base for deriving subject headings.

4) The basic steps in chain indexing are (1) Construction of the class number of the subject of the document; (2) Representation of the Class Number in the form of a Chain; (3) Determination of links like Sought Links (SL), Unsought Links (USL), False Links (FL), and Missing Links (ML); (4) Preparation of Specific Subject Heading; (5) Preparation of Subject Reference Headings; (6) Preparation of Subject Reference Entries; (7) Preparation of Cross References, if any; and (8) Alphabetisation.

5) Principle of context dependency in PRECIS is seen as a combination of context and dependency. When this principle is followed in a PRECIS input string, each term is qualified and sets the next term into its wider context. In other words, the meaning of each term in the string depends upon the meaning of its preceding term and taken together, they all represent the single context. Each term is hence dependent, directly or indirectly, on all the terms which precede it.

6) Two-line-three part entry structure is followed in PRECIS. The first line is occupied by two parts—Lead and Qualifier, which together constitute the Heading. Lead is occupied by the approach term and Qualifier position is occupied by the term(s) that sets the Lead into wider context. The second line is occupied by the third part, i.e. Display position which consists of those additional set of qualifying terms, which rely upon the heading for their context.

   Index entries in PRECIS are basically generated in three formats: (a) Standard format— entries are generated when any of the primary operators (0), (1), and (2) or its dependent elements appear in the Lead; (b) Predicate transformation— entries are generated under a term coded (3) that immediately follows a term coded either by (2) or (s) or (t); and (c) Inverted format— entries are generated whenever a term coded by an operator in the range from (4) to (6) or its dependent elements appear in the Lead.

7) Different operational stages of POPSI may be categorised under three components: Analysis, Synthesis, and Permutation. The work of 'Analysis' and 'Synthesis' is primarily based on the postulates associated with the deep structure of SILs for generating organising classification. The task of analysis and synthesis is largely guided by the following POPSI-table. The work of 'Permutation' is based on cyclic permutation of each term-of-approach, either individually or in association with other terms for generating associative classification effect in alphabetical arrangement.

8) The major steps in formulating index entries according to POPSI are (1) Content Analysis, (2) Formalisation, (3) Standardisation, (4) Modulation, (5) Preparation of Entry for Organising Classification, (6) Approach-term Selection, (7) Preparation of Entries of Associative Classification, and (8) Alphabetisation.

9) In pre-coordinate indexing system, coordination of component terms is carried out before the users come to search the index file, i.e. at the time of indexing by the indexer (i.e. at input stage) in anticipation of the users' approach. But in post-coordinate indexing, component terms of a subject are kept separately uncoordinated by the indexer, and the user does the coordination of concepts at the time of searching (i.e. at the output stage).

10) Search devices used to avoid false coordination of terms in Post coordinate indexing are: (a) Use of Pre-coordinated Terms, (b) Links, (c) Roles, and (d) Weighting.

11) Different varieties of keyword indexing include Key Word in Context (KWIC) Indexing, Key Word Out of Context (KWOC) Indexing, Key-Word Augmented-in-Context (KWAC) Indexing, Double KWIC Indexing, KWWC (Key-Word-With-Context) Indexing, KEYTALPHA (Key-Term Alphabetical) Indexing, KLIC (Key-Letter-In-Context) Indexing, and WADEX (Word and Author Index).

12) Library is a collection of databases. For example, Accession register is a database, Card catalogue is a database, Circulation records constitute a database, Member register is a database. We need to add, edit and maintain all these databases regularly. The advantages of computerised indexing system are as follows: Redundancy can be reduced; Inconsistency can be avoided; Data can be shared; Standards can be enforced; Security restrictions can be applied; Integrity can be maintained and conflicting requirement may be solved.

13) A typical computerised indexing system has four major components: 1) Database, 2) Search process, 3) Language of indexing, and 4) User interface.

    Computerised indexing systems can be categorised into following four groups on the basis of the varieties of information demands of different user groups: 1) Online Database Indexing, 2) Optical Disk based Database Indexing, 3) OPAC based Indexing, and 4) Indexing Internet Resources.

14) A comparative study of online database indexing and Optical Disk based indexing may be done as follows:

Optical Disk based Indexing	Online Database Indexing
1) Locally searchable. No requirement for communication channel.	1) Remotely searchable. Communication network is a must.
2) Cost based procurement.	2) Renewal based procurement.
3) End user searching.	3) Intermediary searching.
4) Offline access. Generally updated on monthly basis.	4) Online access. Updating is daily and sometimes hourly.
5) Contains multimedia database to reference databases.	5) Bibliographic databases those are large and frequently updated.
6) Suitable for homogeneous user environment.	6) Suitable for large number of simultaneous users.

15) Design and development of the computerised indexing system mainly depends on the methods used for organising records in the file. Some of the important logical file organisations are: (1) Sequential File, (2) Inverted File, (3) Indexed Sequential Files, (4) Chained Files, (5) Tree Structured Files, and (6) B-Tree.

16) Different methods of computerised indexing include: (1) Statistical Method, which may consist of (1.1) Term Frequency method, (1.2) Relative Frequency Method, and (1.3) Term weighting method; (2) Linguistic Method, (3) Artificial Intelligence (AI) based Indexing System, which may consist of (2.1) Natural Language Processing (NLP) based Indexing System, and (2.2) Expert System based Indexing System,

17) Indexing Internet resources or Web indexing means the following:

    a) search engine indexing of the Web,

    b) creation of metadata,

    c) organisation of Web links by category, and

    d) creation of a Website index that looks and functions like a back-of-book index.

18) Different types search engines used in indexing Internet resources are: (1) General Search Engine, (2) Regional Search Engine, (3) Meta-search Engine, (4) Subject Specific Search Engine, and (5) Directory-based Search Engines.

19) In order to find information from the Internet, a search engine employs special software robots, called spiders (also called crawler, worm, wanderer, gatherer, etc.), which traverse the web, following links between pages. It builds the list of the words found on Web sites, called Web crawling. Here, the spider begins with a popular site, indexing the words on its pages and following every link found within the site. In this way, the spidering system quickly begins to travel across the Web and takes note of the (a) words within the page, and (b) location of the words - title, subtitles, meta tags and other positions of the Web page. Meta tags allow the owner of a page to specify key words and concepts under which the page will be indexed. The meta tags also guide the search engine in choosing which of the several possible meanings for these words is correct. After completion of finding information on Web pages by the spiders, the search engine stores the information in a way that makes accessible to users. In the simplest case, a search engine just stores the word and the URL where it was found. To make for more useful results, most search engines store more than just the word and URL.

20) The "semantic web" is an approach to extend the web with semantic information to avoid wrong hits. The conception of Semantic Web is characterised by developing tools and technologies like languages, standards and protocols so that the Web becomes meaningful. Technologies involved in the development of the Semantic Web are the Uniform Resource Identifier (URI) for identifying documents uniquely and globally, XML (eXtensible Markup Language) for structuring the data semantically, RDF (Resource Description Framework) to base the structure of the documents on a common model base, Ontologies (to define the objects/entities and the interrelations between these objects/entities), etc.

11.10 KEYWORDS

Action : An elementary category associated with POPSI which refers to an idea denoting the concept of 'doing'. An action may manifest itself as Self Action or External Action.

Assigned Indexing : The process of indexing in which a human indexer selects one or more subject headings or descriptors from a list of controlled vocabulary to represent the subject(s) of a work. Also known as *Assignment Indexing* and *Concept Indexing*.

Associative Classification : It refers to a classification in which a subject is distinguished from all other subjects based on the reference of how it is associated with other subjects, without reference to its COSSCO relationships. The result of associative classification is always a relative index.

Back-of-the-book Index : An index which shows where exactly in the text of a document a particular concept (denoted by a term) is mentioned, referred to, defined or discussed.

Base : It is a particular manifestation or manifestations of a particular elementary category under which all or major portion of related information are brought together.

Boolean Operators : AND, OR, and NOT. Used to combine search terms. AND finds only records that contain both terms. OR finds records that contain either term. NOT finds records that contain the first term but not the second term.

Chain Indexing : The process of deriving subject index entries based on the extracted vocabulary of a notational scheme of classification. It retains all necessary context but removes unnecessary context.

Classaurus : It is an elementary category-based (faceted) systematic scheme of hierarchical classification in verbal plane incorporating all the necessary features of a conventional information retrieval thesaurus. It is used as vocabulary control device in POPSI.

Coextensive Subject Index : A subject index entry, in which a term, phrase, or a set of terms define precisely the full thought content of the document. Here extension and intension of the ideas are equal to the thought content of the document.

Common Modifier : It refers to the name of a place (space), Time, Environment), and Form.

Computerised Indexing : A method of indexing in which an algorithm is applied by a computer to the title and/or text of a work to identify and extract words and phrases representing subjects, for use as headings under which entries are made in the index.

Concrete : An elementary category suggested by Kaiser to refer to things, place and abstract terms, not signifying any action or process.

Content Designation : The act of making a bibliographic record machine readable by encoding its various elements according to a specified scheme.

Core : It is a particular manifestation or manifestations of one or more elementary category under which all or major portion of related information are brought together within a recognised Base.

COSSCO Relationship : It is a relationship in which COordinate—Superordinate—Subordinate—COllateral (COSSCO) relationships of a subject are shown.

Deep Structure of Subject Indexing Languages (DS-SIL) : DS-SIL refers to the logical abstraction of the surface structures of outstanding SILs like Cutter, Dewey, Kaiser and Ranganathan.

Derived Indexing : The process of indexing in which terms to be used to represent the content of the document are derived directly from the document itself. Also known as *Derivative Indexing*.

Discipline : An elementary category associated with POPSI that includes the conventional fields of study, or any aggregate of such fields, or artificially created fields.

Entity : An elementary category associated with POPSI which includes manifestations having perceptual correlates, or only conceptual existence, as contrasted with their properties, and actions performed by them or on them.

False Drops : Retrieval of unwanted documents because of the false coordination of terms at the time of searching.

Input String : A set of terms arranged according to the role operators which act as instructions to the computer for generating index entries.

Item Entry System : A type of post-coordinate indexing system in which It takes the opposite approach to term entry system and prepares a single entry for each

document (item), using a defined physical form, which permits access to the entry from all appropriate headings. Here, items are posted on the term.

Keyword : A term that is chosen, either from the actual text or from the queries of the searcher, that is considered to be a 'key' to finding certain information.

Keyword Indexing : The process of using significant words from a title or an abstract or sometimes from the text of the document as index entries.

KWIC Indexing : Key Word In Context, format for showing index entries within the context in which they occur.

KWOC Indexing : Key Word Out of Context, the use of significant word from titles for subject index entries, each followed by the whole title from which the word was taken.

Meta Search Engine : a program that allows to search across many search engines at once.

Modifier : It refers to a qualifier used to modify any one the elementary categories D, E, A and P associated with POPSI.

Nesting : Grouping terms within parentheses to specify the order in which they will be combined. Terms in the innermost parentheses will be combined and searched first. Without parentheses, terms will be combined in left-to-right order.

Ontology : A formal specification of a representational vocabulary for a shared domain of discourse—definitions of classes, relations, functions, and other objects. Ontologies define data models in terms of classes, subclasses and properties to enhance the functioning of the Web.

Organising Classification : In organising classification compound subjects are based on genus-species, whole-part, and other inter-facet relationships. Organising classification distinguish and rank each subject from all other subjects with reference to its COordinate—Superordinate—Subordinate—COllateral (COSSCO) relationships.

Post-Coordinate Indexing : An indexing model in which terms associated with the content of the document are kept separately in the index file by the indexer and the searcher coordinate coordinates the terms at the time of searching or output stage. Also known as 'coordinate indexing'.

PRECIS : PREserved Context Index System, a subject indexing technique in which an open-ended vocabulary can be organised according to a scheme of role operators, usually for computer manipulation.

Pre-Coordinate Indexing : An indexing model in which terms associated with the content of the document are coordinated by the indexer by following the syntactical rules of given indexing language at the time of indexing or input stage for use in the retrieval of information collection on compound and /or complex concepts.

Process : 'An elementary category suggested by Kaiser to refer to mode of treatment of the subject by the author, an action or process described in the document, and an adjective related to the concrete as component of the subject.

Property : An elementary category associated with POPSI which refers to the idea denoting 'attribute'.

Role Operator : Role operators consist of a set of alpha-numeric notations which specifies the grammatical role or the function of the indexed term and regulates the order of terms in the input string. Role operators and their associated rules also serve as the computer instruction for determining the format, typography and punctuation associated with each index entry.

Search Engine : A retrieval tool on the World Wide Web that, in general, matches keywords input by a user to words found at websites. The more sophisticated search engines may allow other than keyword searching.

Semantic Web : The "semantic web" is an approach to extend the web with semantic information to avoid wrong hits by developing tools and technologies like languages, standards and protocols so that the Web becomes meaningful.

Special Modifiers : A special modifier refers to a qualifier which is used to qualify/modify only one of the elementary categories associated with POPSI.

Stop-word List : The 'stop-word' list refers to a list of words, which 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'

Subject Analysis : The process of identifying the different component

ideas associated with the thought content of the document and establishing the interrelationships among those component ideas.

Subject Gateways : Organized lists of web pages, divided into subject areas by human indexers.

Subject Heading : A word or group of words representing the subject of a document.

Subject Index : A tool that exhibits the analysed contents of the collection of documents (either in the library or database).

Subject Indexing : The process of representing the informational content of the document by analysing its content and translating the result of analysis into an indexing language for creating a surrogate record for it, especially subject access points, in an index.

Term Entry System : A type of post coordinate indexing system in which index entries for a document are made under each of the component terms associated with the thought content of the document. Here, terms are posted on the item.

Web Indexing : Web indexing means providing access points for online information materials, which are available through the use of World Wide Web browsing Software.

World Wide Web (WWW) : a network of many thousands of servers linked together by a common protocol.

11.11 REFERENCES AND FURTHER READING

Austin, Derek. *PRECIS: A Manual of Concept Analysis and Subject Indexing*. 2nd ed. London: British Library Bibliographic Services Division, 1984. Print.

Chakraborty, A. R. and Bhubaneswar Chakraborty. *Indexing: Principles, Processes and Products.* Calcutta: World Press, 1984. Print.

Chowdhury, G. G. *Introduction to Modern Information Retrieval.* 2nd ed. London: Facet Publishing, 2004. Print.

Dykstra, Mary. *PRECIS: A Primer*. Rev. Reprint. Metuchen, NJ & London: Scarecrow Press, 1987. Print.

Foskett, A. C. *Subject Approach to Information*. 5th Ed. London: The Library Association, 1996. Print.

Ghosh, S. B. and J. N Satpathi., *Eds. Subject Indexing Systems: Concepts, Methods and Techniques*. Calcutta. IASLIC, 1998. Print.

Guha, B. *Documentation and Information: Services, Techniques and Systems*. 2nd rev ed. Calcutta: World Press, 1983. Print.

International Standards Organization. *Documentation—Methods for Examining Documents, Determining their Subjects, and Selecting Indexing Terms,* 1985. Print.

Jones, Karen Sparck, Ed. *Information Retrieval Experiment*. London: Butterworth, 1981. Print.

Lancaster, F. W. "Do Indexing and Abstracting have a Future"? *Anales de Documentation*, no 6, 2003, 137-144. Print.

Olson, Hope A. *Subject Analysis in Online Catalogs*. 2nd edition. Englewood: Libraries Unlimited, 2001. Print.

Sarkhel, Juran Krishna. *Information Analysis in Theory and Practice*. Kolkata: Classique Books, 2001. Print.

Sarkhel, Juran Krishna. "Subject Indexing, Vocabulary Control and Recent Developments in Cataloguing". *Library Cataloguing Theory*. (BLIS 04)New Delhi, Indira Gandhi National Open University, 1999. 180p. Print.

Taylor, Arlene G. *Introduction to Cataloguing and Classification*. 10th edition. New Delhi: Atlantic Publishers, 2007. Print.

Wellisch, H. H. *Indexing from A to Z.* 2nd. ed. New York: H. W. Wilson, 1995. Print.

Wilson, T. D. *An Introduction to Chain Indexing*. Hamden, Conn.: Linnet Books, 1971. (Programmed texts in library and information science). Print.