UNIT 4 INDEXING SYSTEMS AND TECHNIQUES

Structure

4.0 OBJECTIVES

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

After reading this Unit, you will be able to:

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

4.1 INTRODUCTION

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

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

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

4.2 INDEXING PRINCIPLES AND PROCESS

4.2.1 Purpose of Indexing

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

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

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

4.2.2 Problems in Indexing

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

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

b) Multidimensional users need for information;

c) Choice of terms from several synonyms;

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

e) Distinguishing homographs;

f) Identifying term relationships — Syntactic and Semantic;

g) Depth of indexing (exhaustivity);

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

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

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

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

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

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

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

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

4.2.3 Indexing Process

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

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

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

c) What is the nature of collection?

d) What is the characteristics of user population?

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

f) Display or physical appearance of the index.

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

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

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

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

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

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

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

ii) **Identification of concepts**

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

iii) **Selection of concepts**

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

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

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

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

4.2.4 Indexing Language

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

4.2.5 Theory of Indexing

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.2.6 Indexing Criteria

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

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

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

4.2.7 Indexing Policy: Exhaustivity and Specificity

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

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

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

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

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

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

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

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

4.2.8 Quality Control in Indexing

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

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

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

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

Translation failures may be of three types:

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

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

c) Omission of important term(s).

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

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

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

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

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

Self Check Exercises

1) Discuss the processes of subject indexing.

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

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

.....

.....

.....

.....

.....

.....

4.3 PRE-COORDINATE INDEXING SYSTEMS

4.3.1 Cutter's Rules for Dictionary Catalogue

It was Charles Ammi Cutter who first gave a generalised set of rules for subject indexing in his *Rules for a Dictionary Catalogue (RDC)* published in 1876. Cutter never used the term 'indexing'; rather he used the term 'cataloguing'.

Cutter identified two problems in organizing library materials in respect of developing subject headings: (a) problem relating to specific subject heading, and (b) order of terms in compound subject heading. Representation of the specific subject of a document by a single word or phrase appeared to be difficult in most cases. So, the complete analysis of the subject of the document is made and this situation paved the way for subject indexing from subject cataloguing. In order to solve the problems of constructing subject headings, Cutter, in his *RDC*, provided rules for specific as well as compound subject headings.

Specific Subject Heading

Cutter did not define the specific subject heading, but the idea was evident in Rule 161 of RDC. In regard to specific subject heading, Cutter advocated, *"Enter a work under its subject heading, not under the heading of a class which includes that subject"[Rule 161].* According to this rule, 'Roses' should be entered under the 'Roses' but not under the general category 'Flower'. However, Cutter gave the following example as a specific subject heading:

Subject of the Work	***Specific Subject Heading***
Movement of fluids in plant	Botany, Physiological

Here the work is entered under the heading of the class, which includes that subject, i.e. under broader class. Thus we see that in his own example Cutter could not maintain uniformity. Such inconsistency may be attributed to the fact that:

a) Cutter envisaged a set of 'stock subjects' under which every book had to be accommodated; and

b) Cutter could not visualize the subject which had no names or not nameable.

This suggests that though Cutter provided new idea for subject cataloguing but his examples were not commensurate with his ideas. There is a great difference between today's 'specific subject' and Cutter's 'stock subjects'. We should not forget the period of Cutter; the development of different subjects and multiple relationships between subjects which basically took place after the Second World

War. It was rather impossible to visualise and compile a list of 'stock subjects' to represent the subject of each and every document that is produced today.

Compound Subject Heading

Rule 175 of *RDC* states: *"Enter a compound subject name by its first word inverting the phrases only when some other word is decidedly more significant or is often used alone with the same meaning as the whole name".*

According to this rule, order of the component terms in compound subject heading should be the one that is decidedly more significant. But Cutter could not prescribe how one will come forward to decide which one is more significant. The question of significance varies from user to user. For example, a document on 'Social psychology' may be approached both by Sociologists and Psychologists. If this rule is followed, there would be two subject headings for the same document: (1) Psychology, Social; and (2) Social Psychology. Thus, we see that the decision in respect of 'significant' is left to the judgment of individual indexer, which is subjective one.

Cutter also provided some specific rules and guidelines in respect of the followings:

a) **Person vs. Country**

Entry will be under person in case of single/personal biography. Entry will be under country in case of history, event, etc.

b) **Country vs. Event**

Entry under event if it is proper noun. Entry under country if it is common noun e.g.
Freedom fighters in India: India, Freedom Fighters

c) **Subject vs. Country**

In scientific subjects, entry will be under the subject qualified by place:

Geology, India.

In areas, such as History, Government, Commerce, entry will be under place qualified by subject:

India—British Period.

For Humanities, Literature, Arts, etc., adjectival form is to be used:

Indian Painting.

d) **Between overlapping subjects,** entry will be according to the importance of the subject;

e) **Choice between different names**

- *Language*—If there are two languages out of which one is English, entry will be under English;
- *Synonyms*—Entry will be under one word with reference to the others;
- *Antonyms*—Entry will be under one word with reference to others;

f) **Compound Subject Heading*:*** The following examples show how Cutter tried to make a solution regarding compound subject heading. Each of these examples may have form and period subdivisions.

- A Noun preceded by an adjective, e.g. ***Political Economy, Ancient History;***
- A Noun preceded by another noun used as an adjective, e.g. ***Death Penalty, Flower Fertilization;***

- A Noun connected to another noun with a preposition—the direct form will be used, e.g. ***Patient with heart disease, Fertilization of flowers, Penalty of death;***
- Phrases or Sentence — Direct forms will be used, e.g. ***Medicine as Profession.***

g) **System of References** — Cutter advocated the following rules:

Rule 187: Make references from general subjects to their various subordinate subjects, and also to coordinate and illustrative subjects.

Rule 188: Make references occasionally from specific to general subject.

Cutter's system of references was from general to specific generally and specific to general occasionally. For the purpose of selecting subordinates for referencing, Cutter himself has prescribed the use of Dewey Decimal Classification and Expansive Classification.

Though Cutter tried to solve the problems of subject cataloguing of documents as it was existed, but he was not successful totally as he could not accommodate all the subjects in a single list and to make a uniform standard rules about the order of precedence of multi-worded subjects expressively.

4.3.2 Kaiser's Systematic Indexing

Julius Otto Kaiser systematised alphabetical subject heading practice by developing the principles behind Cutter's rules so as to form consistent grammar logic. Kaiser was the first person who applied the idea of Cutter in indexing micro documents in the library of Tariff Commission as its librarian. He started from the point where Cutter left. J. Kaiser in his "Systematic Indexing", published in 1911, pointed out that compound subjects might be analysed by determining the relative significance of the different component terms of compound subject through classificatory approach. He categorized the component terms into two fundamental categories: (1) Concrete and (2) Process. According to Kaiser,

Concrete refers to

- Things, place and abstract terms, not signifying any action or process; e.g. gold, India, Physics, etc.

Process refers to

- Mode of treatment of the subject by the author; e.g. *Evaluation* of IR system, *Critical analysis* of a drama.
- An action or process described in the document; e.g. *Indexing* of web documents.
- An adjective related to the concrete as component of the subject; e.g. Strength of metal.

Kaiser suggested a rule that a '*process'* should follow *'concrete'*. According to Kaiser, the '*concrete'* and *'process'* for the subject of a document on 'Heat treatment of metals' would be 'metals' and 'heat treatment' respectively and its subject heading would be: METALS—Heat Treatment.

Kaiser also envisaged that the 'concrete' is the main term, the 'process' gives the action state of the 'concrete', and the country supplies the locality where the action takes place. In the case of no 'concrete' and no 'place', the 'process' remains. But it is preferable to complete the subject heading with 'concrete' terms in all cases. As a result, we find the following unnatural subject headings:

Documents on	*Subject Headings*
Painting	PAINTS—Application
Painting of cars	CARS—Painting
Wage	LABOUR—Price
Strike	LABOUR—withdrawal

From the above examples it appears that Kaiser could not maintain consistency in the formulation of subject headings. He gave the idea of conceptual categorization but could not apply it properly in all cases. Beside this, Kaiser laid a rule that if a subject deals with place, double entry (*'Concrete—Place—Process'* and *'Place—Concrete—Process'*) is to be made. For example, index entries for 'Production of jute in West Bengal' would be: ***JUTE—West Bengal—Production; WEST BENGAL—Jute—Production.***

To help an enquirer in getting to the related topics, Kaiser recommended an elaborate system of references. He recommended that every 'concrete' term should be equipped with cross-references to both hierarchically superiors and inferiors. However, he did not recommend cross-references from 'process' terms.

In short, Kaiser's achievements regarding subject indexing are:

- *categorization* of composite terms through classificatory approach for the first time;
- a *general rule of order of precedence*, i.e. the 'process' term should follow the 'concrete' term;
- *definition* of those terms, of which 'process' is identified properly, that is, he gave the characteristics of 'process' by which it can be identified properly;
- *double entry system* for a subject dealing with place/locality; and
- elaborate *system of references*.

Though Kaiser brought much development in subject indexing, it had some drawbacks. These are:

a) There is no provision for entry under the 'process' term and as a result it fails to satisfy the users' approach by the 'process' term;

b) Double entry system recommended by him is uneconomical;

c) Kaiser's prescription of analysis of subject into C*oncrete* and *Process* sometimes leads to unnatural headings; and

d) Kaiser left out the concept of 'time'.

Though Cutter and Kaiser could not solve many of the problems of subject indexing, still they could set the ball rolling to progress further to find out a logical solution to the problem of subject indexing.

4.3.3 Chain Indexing

Dr. S. R. Ranganathan developed a method of indexing, called *chain procedure* of subject indexing or simply Chain Indexing. It is a method of deriving alphabetical subject entries from the chain of successive subdivisions of subjects needed to be indexed leading from general to specific level. According to Ranganathan, chain indexing is a "procedure for deriving class index entry (i.e. subject index entry) which refers from a class to its class number in a more or less mechanical way." A note is also given with the above definition as:

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

The term 'chain' refers to a modulated sequence of subclasses or isolates. Since the chain expressed the modulated sequence more effectively in a notational classification of subjects, this method takes the class number of the document concerned as the base for deriving subject headings not only for specific subject entry but also for subject reference entries. Therefore, the chain is nearly derived from a classification scheme, but not necessarily from a scheme. The method is intended to offer general as well as specific information to all information seekers by deriving subject headings from the chain of successive subdivisions that leads from the general to most specific level. While classifying a subject of a document we follow the same steps as those required for subject indexing in verbal plane (i.e. in natural language), the only difference is that in a classification scheme instead of representing the subject in verbal plane we use an artificial language to represent the subject in notational plane. The nature and structure of the classification scheme controls the structure of the subject headings drawn according to the chain procedure. The basic steps in chain indexing are enumerated below:

a) Classify the subject of a document by following a preferred classification scheme.

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

c) Determine different kinds of links: sought, unsought, false and missing links. Sought links denote the concepts (at any given stage of the chain) that the user is likely to access; unsought links are those that are not likely to be used as access points; false links are those that really do not represent any valid concept, mostly these are connecting symbols. Missing links represent those concepts that are not available in the preferred classification scheme, these are inserted by the indexer by means of verbal extension at the chain-with-gap corresponding to the missing isolate in the chain whenever there is such a need.

d) Derive specific subject heading for the specific subject entry from the last sought link and moving upwards by taking the necessary and sufficient sought links in a reverse rendering or backward rendering process. If the subject includes a space or a time isolate or a form, then break the chain into parts of the point(s) when the digit denoting a space or a time or a form. In such a situation specific subject heading is to be derived from last sought link of each part in reverse rendering process and during the reverse rendering process, the chain just before the part denoting time, space or form must be inverted, and the rest of the chain appended to it.

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

f) Prepare subject references or '*See also*' references from each subject reference heading to its specific subject heading when a subject heading starts from last sought link denoting space rendered, or a time or a form, prepare '*See*' references instead of '*See also*' references from subject headings to specific subject heading.

g) Prepare '*see*' references for each alternative and synonymous term used in the specific as well as subject reference headings.

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

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

1) *Subject Statement:* Researches in the promotion of student welfare programme through psychological counseling in Indian schools.

2) *Class no.:* 371.713072054 [according to DDC, 21st Edition]

3) *Representation of the class number in the form of a chain and determination of different kinds of links:*

300	Social Sciences
370	Education [SL]
371	Schools [SL] and their activities; Special education
371.7	Student welfare [SL]
371.71	Student health [SL] and related topics
371.713	Mental health services [SL]
371.7130	[FL]
371.71307	Study and Teaching [USL]
371.713072	Research [SL]
371.7130720	[FL]
371.71307205	Asia [USL]
371.713072054	India [SL]

4) *Preparation of specific subject heading, subject reference headings and cross references:*

Specific Subject Heading:

Research, Mental health services, Student welfare, Schools, India,

Subject Reference Headings:

Research, Mental health services, Student welfare, Schools

Mental health services, Student welfare, Schools

Student health, Student welfare, Schools

Student welfare, Schools

Schools, Education

Education

Cross References:

India, Research, Mental health services, Student welfare, Schools

See

Psychological counseling

Research, Mental health services, Student welfare, Schools, India

5) *Preparation of Index Entries*

Research, Mental health services, Student welfare, Schools, India
371.71307205

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

Research, Mental health services, Student welfare, Schools 371.713072

See also

Research, Mental health services, Student welfare, Schools, India

Mental health services, Student welfare, Schools 371.713

See also

Research, Mental health services, Student welfare, Schools, India

Student health, Student welfare, Schools 371.71

See also

Research, Mental health services, Student welfare, Schools, India

Student welfare, Schools 371.7

See also

Research, Mental health services, Student welfare, Schools, India

Schools, Education 371

See also

Research, Mental health services, Student welfare, Schools, India

Education 370

See also

Research, Mental health services, Student welfare, Schools, India

Psychological counseling *see* Mental health services

India, Research, Mental health services, Student welfare, Schools

See

Research, Mental health services, Student welfare, Schools 371.71302

6) *Alphabetisation*

Arrange the above entries according to single alphabetical order.

Chain indexing, a variety of pre-coordinate indexing, has some advantages like it is more or less a mechanical system and it is economic also. The greatest criticism against this method had been that only the last link provides the specific heading. Other entries are more general successively. Too much dependence upon classification scheme is another criticism. Ranganathan in a later paper clarified that index headings may be drawn both by forward rendering and backward rendering method. If we look deeply into the concept of Chain Indexing, it should be clear that Chain Indexing need not be dependent upon a particular classification scheme but it is based on classification principle and method.

4.3.4 Relational Indexing

J. E. L. Farradane devised a scheme of pre-coordinate indexing known as *Relational Indexing*. This indexing system was developed first in the early 1950s and has been modified since then. The latest change may be noted from Farradane's own paper in 1980. The basic principle of Farradane's Relational Indexing is to identify the relationship between each pair of terms of a given subject and to represent those relations by relational operators suggested by him and thus creating 'Analets'.

Relational operators are special symbols which link the isolates to show how they are related and each operator is denoted by a slash and a special symbol

having unique meaning. That is why the system propounded by Farradane is known as the *System of Relational Analysis*. He has formulated the basic idea of relational analysis by studying the development of learning process with particular reference to child psychology. Two or more isolates linked by *relational operators* as a subject statement constitute *Analet*.

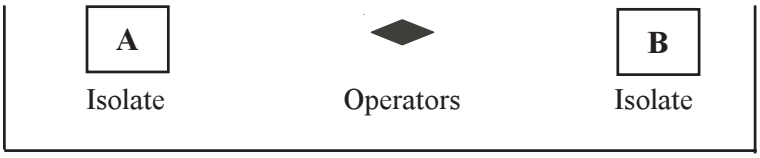


Fig. 4.1: Analet

Two types of relationship has been identified and they are based on the qualities of experience of how children learn by developing power of discrimination in time and space and stages of discrimination in each of the following areas:

- The time sense involving three stage of discrimination towards complete association.
- Awareness of degree of distinctness (in space) involving three stages towards discrimination between terms.

In time, the first stage is 'non-time'—the co-occurrence of two ideas without reference of time; the second stage is 'temporary'—the co-occurrence of two ideas from time to time, but not permanently; and the third is 'fixed'—the permanent co-occurrence of two ideas. In space, the stages of discrimination are: the first is 'concurrent'— two concepts which are hard to distinguish i.e. parallel; the second is 'non-distinct'— two concepts which have much in common, e.g. cities and villages, science of religion; the third is 'distinct'—two concepts which can be completely distinguished.

Combination of these two mechanisms leads to nine categories of relations, which are shown through the following matrix:

Associative Mechanism (Increasing Association)

Discriminatory Mechanism (Increasing clarity of perception or recognition)

	Awareness (Non-time)	**Temporary Association**	**Fixed Association**
Concurrent Conceptualization	Concurrence / o	Self-activity / *	Association / ;
Non-distinct Conceptualization	Equivalence /=	Dimensional / +	Appurtenance / (
Distinct Conceptualization	Distinctness /)	Reaction / Action / ¾	Functional Dependence (Causation) /:

Fig. 4.2: Farradane's Relational Operators

Table 4.1: Farradane's Relational Operators with Illustrative Examples

Relational Categories and their operators	When to Use	Subject	Analet
1. Concurrence / 0	i) Mere co-existence of two terms like A in the presence of B; ii) Bibliographic form	Dictionary of Chemistry	Chemistry / 0 Dictionary
2. Self-activity / *	i) Intransitive verb; ii) Relation between the agent and activity	i) Migration of birds; ii) Broadcasting through satellite;	i) Bird / * Migration ii)Satellite /* Broadcasting
3. Association / ;	i) Unspecified association; ii) Relation of an agent, tool, thing/ application, discipline, indirect or calculated property, part or method of an action.	i) Application of computer in indexing; ii) Efficiency of calculator.	i) Indexing /; Computer ii) Calculator /; Efficiency
4. Equivalence / =	Equivalence of two things like synonyms, quasi-synonyms, something considered as something else, etc.	i) Vegetables like tomato; ii) Phosphate used as a fertilizer	i) Tomato / = Vegetables ii) Phosphate / = Fertilizer
5. Dimensional / +	Expresses position in space and time.	Petrochemicals plants at Haldia	Petrochemical Plants / + Haldia
6. Appurtenance / (	i) Whole-part relation; ii) Physical property belonging to a substance.	i) Floor of the building; ii) Density of acid;	i) Building / (Floor ii) Acid / (Density
7. Distinctness /)	Expresses the idea of an alternative or substitute	i) Photograph of tiger; ii) Gold and silver	i) Tiger /) Photograph ii) Gold /) Silver
8. Reaction/Action / –	Expresses action or effect of a thing or process on another thing or process	Purifying water	Water / – Purifying
9. Functional Dependence / :	i) Expresses the relation of one thing causing or producing something; ii) Product from a raw material or process.	i) Cyclone due to depression; ii) Extraction of minerals from ore	i) Depression / : Cyclone ii) Ore / : Minerals / – Extraction

Farradane's Relational Indexing offers no rule for filing order for the component terms in an index entry. If the reverse order of the components is required then the necessary operators are also required to be written in the reverse order. For example,

'Depression / : Cyclone' can also be written as *'Cyclone : / Depression'*.

Farradane has claimed that his system gives very good results in terms of recall and relevance, and that once the basic theory underlying the system is grasped, it can be applied very easily; it takes very little time to construct the *analet* once the subject of the document is determined. This claim is not borne out by other workers in the field. It was found that even after a fair amount of experience, the amount of time taken to index a document using this system was such as to make it highly uneconomic, while the results obtained were no better than with other systems. Farradane's Relational Indexing has been the subject of scholarly research, but was not implemented widely. Still, we can say that Farradane's marked improvement in the area of subject indexing was:

- analysis of relationship among terms;
- use of relational operators; and
- one to one relationship among analets.

4.3.5 Coates's Subject Indexing

Idea of E. J. Coates is not considered as original in nature. From the contributions of Cutter, Kaiser and Ranganathan, the concept of *Term Significance* was drawn. From the contribution of Farradane, the concept of *Term relationship* was drawn. Coates, in his contribution, has made a synthesis of above two concepts. It was advantageous for Coates to apply his idea on British Technology Index (now Current Technology Index) of which he had been the editor from its inception in 1963 until his retirement in 1976.

Term Significance

The most significant term in a compound subject heading is the one that is most readily available to the memory of the enquirer. From this, Coates has developed the idea of *Thing* and *Action* like Kaiser's Concrete and Process.

A *'Thing'* is whatever one can think, that is to say whatever can be thought as a static image. It is the image that comes straight into our mind, i.e. which we can visualize first. It includes not only the names of physical objects but systems and organisations of a mental kind, e.g. Democracy [system].

An *Action* refers to any thing in action or process denotes by term / word.

Example: Heat treatment of aluminium

ALUMINIUM [Action] / Heat treatment [Thing]

Among 'Thing' and 'Action', thing is more significant than Action. So the subject heading would be: Thing / Action. Coates also added the idea of M*aterial*. A *Material* is a state images produced by names of Things and names of Material. A Thing is made up of certain Material and the Material follows the Thing. A Thing has a boundary (i.e. shape) but the Material has none and for this reason the name of the Material is of low significance than the name of a Thing. On the other hand, the image of a Material is made of some static seeming properties—such as, colour, hardness, smoothness, etc. (i.e. the image of a material is produced due to its colour, hardness, smoothness, etc.) and is rated higher in the significance than the name of an *action.* Therefore, the significant order is *Thing—Material—Action* like Personality— Matter—Energy of Ranganathan. *Part* is a component of a thing. Hence, it depends on the thing to which they belong and, thus, giving us the order of significance: *Thing–Part–Material–Action.*

Term Relationship

The order of significance is fundamental in indexing. But it does not answer all questions of the order of the component terms. Apart from significance order, compound subjects may include two or more equally concrete things, or a complex subject may incorporate two or more phrases, which sometimes bring problems in indexing. These problems can be solved if the relationship between components of the compound term is considered. Establishing significance order for a subject heading requires a great deal of intellectual effort.

In order to decide the problems of above mentioned kind, it is necessary to consider:

a) How far relationships between components lead to modifications of the significance formula: *Thing – Material – Action*; and

b) How far the significance formula lead to modification of natural language order of the whole components.

Coates has provided a very valuable corollary of his ideas on significance order. Coates broke each phrase into terms with necessary preposition in between them which he called *Amplified phrase order*. Amplified phrase order is the order of component terms achieved by using the necessary prepositions in between them.

Examples:

Phrase / Compound Subject	***Amplified Phrase order***	***Subject Headings***
Conveyor belt	Belt of conveyor	Conveyor, Belt
Belt conveyor	Conveyor with belt	Conveyor, Belt

In a simple way, we can say that a prepositional phrase in the form of *Action–Preposition–Thing* is used while expressing the idea of a *Thing* being acted upon by an *Action*. Then we have to reverse the phrase omitting the preposition to generate a subject heading in accordance with the *Thing–Action* significance formula as shown in the example: ALUMINIUM / Heat treatment.

Coates has identified a number of relationships by means of a number of prepositions—such as, 'Of', 'For', 'Against', 'With' and 'By'. Coates distinguished 20 different kinds of relationships including *Thing – Action*, and tabulates these to show their relation to the corresponding prepositional phrase. Coates also suggested some rules in respect of order of precedence or order of significance.

Categories of Relationships

Categories of relationships proposed by Coates fall into following two broad divisions:

a) Relationships dealing with properties, actions, materials, viewpoints and parts in relation to one another and to Things [Categories 1—7].

b) Relationships dealing not with actions, parts, and materials as such but with Things differentiated by reference to parts, materials, principles of action, and causative factors [Categories 8—20].

Each category is illustrated below with examples:

Category 1: Action on Thing
Example: Planning of Villages
Subject Heading: VILLAGES [Thing], Planning [Action]
Significance formula: Agrees
Amplified phrase order: Reverses

Category 2: Action on Material
Example: Moulding of plastic
Subject Heading: PLASTIC [Material], Moulding [Action]
Significance formula: Agrees
Amplified phrase order: Reverses

Category 3: Action A on Action B
Example: Corruption in politics
Subject Heading: POLITICS [Action B], Corruption [Action A]
Significance formula: Nil (since no 'Thing' is present)
Amplified phrase order: Reverses

Category 4: Material of Thing
Example: Metal of container
Subject Heading: CONTAINERS [Thing], Metal [Material]
Significance formula: Agrees
Amplified phrase order: Reverses

Category 5: Part of Thing

Example: Skin of animals

Subject Heading: ANIMALS [Thing] Skin [Part]

Significance formula: Nil

Amplified phrase order: Reverses

Category 6: Property of Thing, Material and Action

Examples: i) Intelligence of man

ii) Hardness of iron

iii) Velocity of flow

Subject Headings: i) MAN [Thing], Intelligence [Property]

ii) IRON [Material], Hardness [Property]

iii) FLOW [Action], Velocity [Property]

Significance formula: Nil

Amplified phrase order: Reverses

Category 7: Partial viewpoint on Thing, Material or Action or Property.

Examples: i) Physics of metals (*Partial viewpoint on Thing*)

ii) Ethics of hunting (Partial viewpoint on Action)

Subject Headings: i) METAL [Thing], Physics [Viewpoint]

ii) HUNTING [Action], Ethics [Viewpoint]

Significance formula: Nil

Amplified phrase order: Reverses

Category 8: Things distinguished by citation of Principle of Action

Example: Bridges based on suspension

Subject Headings: BRIDGES [Thing], Suspension [Action]

Significance formula: Agrees

Amplified phrase order: Agrees

Category 9: Things distinguished by citation of Materials

Example: Bricks made of gold

Subject Heading: BRICKS [Thing,], Gold [Material]

Significance formula: Agrees

Amplified phrase order: Agrees

Category 10: Things distinguished by Part.

Example: Books with illustration

Subject Heading: BOOKS [Thing], Illustration [Part]

Significance formula: Agrees

Amplified phrase order: Agrees

Category 11: Thing distinguished by Material or Form of entry which it utilises.

Example: Engines operated by steam

Subject Heading: STEAM [Material] ENGINES [Thing]

Significance formula: Reverse

Amplified phrase order: Reverse

Category 12: Action A distinguished by citation of contributing or underlying Action B

Example: Welding by resistance

WELDING [Action A], Resistance [Action B]

Significance formula: Nil (Because both of them are Action)

Amplified phrase order: Agrees

Category 13: Thing A, serving or supplying or aiming at Thing B

Example: Transistors for radio

Subject Heading: RADIO [Thing B], Transistors [Thing A]

Significance formula: Nil (Because both of them are Thing)

Amplified phrase order: Reverse

Category 14: Thing A or Action A distinguished from homonyms by the fact that it serves Thing B or Action

Example: Accelerators for composting

Subject Heading: ACCELERATORS [Thing A], Composting [Thing B]

Significance formula: Nil (Because both of them are Thing)

Amplified phrase order: Agrees

Category 15: Thing serving or instrumental to Action (i.e. one Thing is used as instrument for action)

Example: Machines for washing

Subject Heading: WASHING [Action] MACHINES [Thing]

Significance formula: Reverse

Amplified phrase order: Reverse

Category 16: Thing A caused by Thing B

Example: Craters caused by meteorites

Subject Heading: METEORITES [Thing B], Craters [Thing A]

Significance formula: Nil

Amplified phrase order: Reverse

Category 17: Thing caused by Action

Example: Crystals caused by condensation

Subject Heading: CRYSTALS [Thing], Condensation [Action]

Significance formula: Agrees

Amplified phrase order: Agrees

Category 18: Action caused by Thing

Example: Reaction caused by enzymes

Subject Heading: ENZYMES [Thing], Reaction [Action]

Significance formula: Agrees

Amplified phrase order: Reverse

Category 19: Action A caused by Action B

Example: Rain caused by depression

Subject Heading: RAIN [Action A], Depression [Action B]

Significance formula: Nil

Amplified phrase order: Agrees

Category 20: Thing or Action at a type of location

Example: Plants found in Water

Subject Heading: WATER [Location], Plants [Thing]

Significance formula: Nil

4.3.6 PRECIS

Dereck Austin developed PRECIS, the PREserved Context Index System, in 1974 as an alternative procedure for deriving subject headings and generating index entries for British National Bibliography (BNB) which since 1952, was following Chain Indexing. Two most important factors played significant role in looking for an alternative method, ultimately resulted in the development of PRECIS: i) ideas of replacing chain indexing technique of BNB; and ii) the decision of the British Library to generate computer produced BNB with all the indexes. While developing PRECIS the following factors were taken into consideration.

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

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

c) Each of the entries should be expressive;

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

e) The system must have sufficient references for semantically related terms.

In the light of the experiences gained through the application of PRECIS in UK, USA, Canada, Australia, Germany and in counties as linguistically and culturally far apart, as Denmark, Italy, Poland, China, etc., Dr. Dereck Austin brought out its second edition in 1984.

Syntax and Semantics of PRECIS

PRECIS consist of two inter-related sets of working procedures: *Syntactical* and *Semantic*. Syntactical relationships in PRECIS are handled by means of a set of logical rules and a *schema of role operators* and *codes*. They refer to the organisation of terms in input strings and their manipulation to generate index entries.

Semantic, i.e. a *priori* relationship between indexing terms and their synonyms are regulated by machine-held thesaurus that serves as the sources of *see* and *see also* references in the index. A thesaurus is generated simultaneously with the preparation of input string. The PRECIS manual goes into great and technical details about the online construction of input record for establishing semantic network of terms.

Principles of PRECIS

The PRECIS is based on two principles:

a) *Principle of Context Dependency:* By name, it suggests that the PRECIS is based on the *Principle of Context Dependency.* Each of the terms in the input string sets the next term into its most obvious wider context (i.e. general to specific). In other words, the meaning of each term in the string depends upon the meaning of its preceding term and taken together, they all represent a single context.

b) *Principle of One-to One Relationship:* When terms are organised in context-dependent order, they form *a one-to one relation.* This simply means that each of the terms in the string is directly related to its next term.

Thus, these two characteristics (*context dependency*, leading to terms linked by *one-to-one relations*) are the basis on which the whole system works and they play a very important role in conveying the meaning of an index entry.

Entry Structure of PRECIS

In order to achieve the principle of context-dependency, *Two-Line-Three-Part* entry structure is followed in PRECIS and this can be demonstrated with the following diagram:

A string of five terms A-B-C-D-E has been represented in one-to-one related sequence. The term A or the term E can act as entry point. But, when C is considered as entry point, the different structure is necessary. A '*Two-Line-Three-Part'* entry structure expresses the relationship of the terms fully (Figure 4.3).

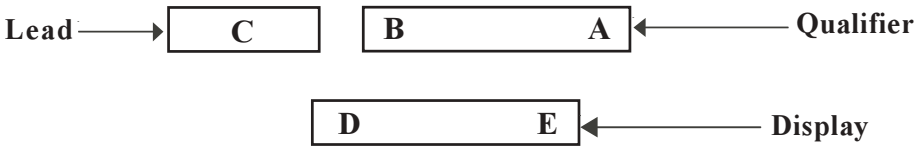


Fig. 4.3 : Entry structure

The first line consists of two units—the *lead* and *qualifier*—is called the *heading,* and the other line is called the *display. Lead* is occupied by approach term or filling term and is offered as the user's access point in the index. *Qualifier* position is occupied by the term(s) that sets the lead into its wider context (i.e. general to specific). Heading consists of terms set down in a narrower-to-wider context order. When *Lead* is occupied by the first term of the input string, the qualifier position is usually kept blank. *Display* position is occupied by those terms that rely upon the heading for their context. Sometimes, the display position becomes empty when lead is occupied by the last term of the input string.

Input String, Role Operators and Codes

It has already been stated earlier that the grammar of PRECIS is based on the codes and role operators, which form the syntactical backbone of PRECIS and regulate the writing of input string. *Input String* refers to a set of terms arranged according to the role operators which act as instructions to the computer. *Role Operators* refer to a set of notations, which specifies the grammatical role or function of the term which follows the operators and which regulates the order of the terms in an input string. The rules associated with role operators serve as computer instructions for generating index entries determines the format, typography and punctuation associated with each index entry.

There are two kinds of Role Operators: *Primary Operators* (earlier called Mainline Operators) and *Secondary Operators* (earlier called Interposed Operators). One of these operators has to be written as a prefix to each of the terms in an input string. *Primary Operators* control the sequence of terms in the input string and also determine the format of entries in the printed index. There are seven primary operators in the range 0 to 6 having built-in filing value for three sets of concepts: (0) is for the 'environment of the core concepts'; (1), (2) and (3) are for the 'core concepts'; and (4), (5) and (6) are for the 'extra-core concepts'. *Secondary Operators* (earlier called 'Interposed Operators') can be introduced into a string at any point to raise its level of exhaustivity, but these operators cannot be used to start a string. These operators are used for three sets of concepts: (f) and (g) are for 'coordinate concepts'; (p), (q) and (r) are for 'dependent elements'; and (s), (t) and (u) are for 'special classes of action'. Any of the secondary operators is always to be preceded by a primary operator to which it relates. The term which has to come in lead position is to be marked by (✓).

An important feature of the revised version of PRECIS is the provision of *Codes* for bringing expressiveness in the resulting index entries. Three types of codes are there – Primary, Secondary, and Typographic codes. There are two sets of primary codes: one set (*theme interlinks*) involves three codes used to link common or related themes in the subject statement, while the other (*term codes*) also involves three codes, used to denote the status of a term in the subject statement. There are also some secondary codes which are divided into two sets: the first set, *differences,* involves three kinds of differences—such as, 'preceding

differences', 'date as a difference' and 'parenthetical differences'. The second set involves *connectives,* which are used to connect any two or more consecutive terms in the input string in order to bring expressiveness in the resulting index entries. There are also some *typographic codes* that are used to bring the desired typographic form of a given term in the resulting index entries.

Schema of Role Operators

Primary Operators

Environment of core concepts	0	Location
Core concepts	1	Key System Thing when action not present. Thing towards which an action is directed, e.g. object of transitive action, performer of intransitive action
	2	Action; Effect of action
	3	Performer of transitive action (Agent, Instrument); Intake; Factor
Extra-core concepts	4	Viewpoint-as-form
	5	Selected Instance: study region, study example, sample population
	6	Form of document; Target user

Secondary Operations

Co-ordinate concepts	f	'Bound' co-ordinate concept
	g	Standard co-ordinate concept
Dependent elements	p	Part; Property
	q	Member of quasi-generic group
	r	Assembly
Special classes of action	s	Roll definer; Directional property
	t	Author-attributed action
	u	Two-way interaction

Codes used in PRECIS Strings

Primary codes

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

Secondary codes

Differences

Preceding differences 1st and 2nd characters:

(3 characters)	$0 Non-lead, space generating
	$1 Non-lead, close-up
	$2 Lead, space generating
	$3 Lead, close-up

3 rd character = number in the range, 1 to 9 indicating level of difference

Date as a difference	$d
Parenthetical difference	$n Non-lead parenthetical difference
	$o Lead parenthetical difference
Connectives	$v Downward reading connective
	$w Upward reading connective

Typographic codes

$e Non-filing part in italic preceded by comma
$f Filing part in italic preceded by comma
$g Filing part in roman, no preceding punctuation
$h Filing part in italic preceded by full point
$I Filing part in italic, no preceding punctuation

Steps in PRECIS

Major steps involved in indexing according to PRECIS include the following:

1) Analysing the thought content of the document concerned and identifying the component terms denoting key concepts.
2) Organizing the component terms into a subject statement based on the principle of context dependency.
3) Determining the role or status of each term in terms of role operators.
4) Assigning the role operators, which signify the syntactical role of each term.
5) Deciding which terms should be the access points and which terms would be in other positions in the index entries, and assigning further codes to achieve these results.
6) Adding further prepositions, auxiliaries or phrases, which would result in clarity and expressiveness in the resulting index entries.
7) Generation of index entries by the computer.
8) Generation of supporting reference entries (i.e. *'see'* and *'see also' references*) from semantically related terms taken from a machine-held thesaurus.

Formats of PRECIS Index

There are three kinds of format in PRECIS: *Standard Format, Inverted Format* and *Predicate Transformation.* Each format is related to one or more role operators (primary) and is used while dealing with the appropriate role operator(s).

Standard Format

Index entries in the *standard format* are generated with the primary operators (0), (1) and (2) through the process, known as *Shunting*, by the computer. Here, a string of terms is marshaled, like railway tracks, in the display position prior to generation of any index entry. As the shunting process begins, the computer will generate the first index entry by pushing the first term of the input string into the lead position. Pushing the second term into the lead position and thereby replacing the existing lead term into the qualifier position will generate the second index entry. The process will continue until the computer reaches the last term of the input string. As soon as any term goes to the lead position, it is printed in bold typeface.

Example: 'Training of labours in India

(0) ³ India

(1) ³ labours

(2) ³ training

India

Labours. Training

Labours. India

Training

Training. Labours. India

Inverted Format

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

Example: 'A bibliography of statistics for librarians'

(1) ³ Statistics

(5) ³ Librarians $01 for

(6) ³ bibliographies

Statistics
– For librarians – Bibliographies

Librarians
Statistics *– For librarians – Bibliographies*

Bibliographies

Statistics *– For librarians*

Predicate Transformation

When an entry is generated under a term coded (3) that immediately follows a term coded either by (2) or (s) or(t)—each of which introduces an action of one kind or another—the *predicate transformation* takes place.

Example: 'In-service training of librarians by teachers

(1) ³ librarians
(2) ³ training $21 In-service $v by $w of
(3) ³ teachers

Librarians
In-service training by teachers

Training. Librarians
In-service training by teachers

Teachers
In service training of librarians

The following example illustrates the input string for a given subject statement and the generation of PRECIS index entries using different operators:

<u>Subject Statement</u>: A report on policies of the Government of India of the applications of computers in manpower planning of universities.

<u>Input String</u>:

(1) ³ universities

(p) ³ manpower

(2) ³ planning

(sub 2 ↑) (2) ³ manpower planning $w of

(s) ³ applications $v of $ in

(3) ³ computers $w for

(s) ³ policies $v of $w on

(sub 2 ↑) (3) ³ Government of India

(1) ³ India

(2) ³ Government

Annotation

1) The term 'applications' coded (s) introduces the idea of tool or instrument, used by an unspecified agent, generally human. This role defining term explains the function of the 'computer' in relation to 'manpower planning'.

2) The term 'policies' refers to an attribute (a property) of the term cited next (i.e. Government of India) and is directed towards the 'applications of computers in manpower planning of universities'.

3) Connectives $v and $w have been used in the string to construct pre-coordinate phrases in the resulting index entries.

4) Upward reading substitutes "Manpower planning' and downward reading substitutes 'Government of India' are preceded by the operators (2) and (3) respectively in order to express the role of the phrase as a whole, since they will affect the position of the substitute in the resulting index entries. While generating index entries, downward reading substitute and upward reading substitute are always ignored when the string is read in upward and downward directions respectively.

5) When the term coded (s) goes to the lead, the term coded (3) will go to the qualifier and the concept directed will go to the display. The extent to which the terms coded (s) justify making as leads depends on the subject field and also on the local practice.

Index Entries:

Universities

Manpower. Planning. Applications of computers.

Policies of Government of India

Manpower. Universities

Planning. Applications of computers. Policies of Government of India

Planning. Manpower. Universities

Applications of computers. Policies of Government of India

Computers

Applications in manpower planning of universities.

Policies of Government of India

Policies. Government of India

On computers for applications in manpower planning of universities

India

Government. Policies on computers for applications in manpower planning of universities

Government. India

Policies on computers for applications in manpower planning of universities

Annotation

In the above examples, Lead and Qualifier are separated by a full stop and two letter spaces. The standard separator between two terms in the entry is full stop

and one space. However, for certain operators, other punctuation's like colon, space, dash, etc. are used. The display is written leaving 2-letter space from the left. For over-run of display, 4-letter space and for over-run of heading 6-letter spaces are left from the margin.

4.3.7 COMPASS

In 1990, it was decided to revise UKMARC and to replace PRECIS by a more simplified system of subject indexing. As a result Computer Aided Subject System (COMPASS) was introduced for BNB from 1991 using the same kind of basic principles of PRECIS and the PRECIS was dropped.

PRECIS was designed for the specific purpose of generating coextensive subject statement at each entry point in a form that was suitable for a printed bibliography. This was not necessarily the best format for online searching. It is worth recapitulating briefly some of the criticism of PRECIS in order to put the COMPASS into proper perspective:

i) The syntactic structure of PRECIS is complex and time consuming. Their complex systems of role operators served to provide the output string for printing, but were not otherwise utilized. Through there is no reason why they should not have been.

ii) PRECIS appears to be imprecise in some aspects; for example, in many instances it does not appear to make any difference whether a concept is coded (1) or (2), which suggests that the operators would not be of much help in searching a computer file, where they might be included.

iii) Place name has been treated in several ways as part of the subject string. Depending on the sense, place name is coded by the operators (0), (1), (5) and occasionally (3).

iv) The author information may be of value, if an individual or a corporate body is closely associated with a particular subject. Persons as subjects, for examples of biographies, also from part of the PRECIS subject string. As a result entries for an individual may be found in both the Author/Title file and the subject file. Common practice for many years has been to file such entries in the Author/Title file, taking this as a name file .If a record is being searched online, it is immaterial whereas in the records a piece of information occurs, so long as it is found if it is there.

v) PRECIS allows very long headings. For example,

Acquisition. Books. Stock. Libraries. Universities. United States
Selection. Approach plans – *Reports*

In this example, everything before "selection" is the heading. Long heading like this are not likely to be shared by more than one index element and the main purpose of distinguishing headings from subheadings seems to be thwarted. Even when more than 100 index elements begin with "Acquisition", a PRECIS index display will repeat this term each time if the other component terms of the heading are different. By contrast, in a system in which the lead term alone always forms the heading, the lead term "Acquisition" could be displayed once for a number of the index elements.

vi) PRECIS index generation rules are quite complex. It is practically difficult for an indexer to keep nearly 200 rules in mind every moment.

vii) User of the PRECIS manual (be he a student, teacher, or a practicing indexer) is too often confronted by the fine distinctions and interpretations, which sometimes seem incomprehensible and, thus, leads to consistency.

Syntactical and Semantic aspects of Compass

Role operators used in COMPASS are similar to PRECIS role operators. In order to minimize the complexity of PRECIS role operators, primary operator's (1), (2) and (3) are used for COMPASS along with the secondary operators' (p), (q), and (r). The PRECIS primary operators' (0), (4), (5) and (6) are not used for COMPASS. Codes in respect of "connectives" –that is, $v and $w are used in the input string to construct pre-coordinate phrases and thus to disambiguate the resulting index entries. In a subject "Assessment of students by teachers", for example, the following index entry:

Students
Assessment. Teachers

may lead to represent another subject "Assessment of teacher by students". In such cases, use of "connectives" with the term (i.e. Assessment) dealing with the concept of actions in the input string will reveal the subject properly from the resulting index entry as displayed below:

Students
Assessment by teacher

Pre-coordination also has a role to play in an on-line information retrieval system. An on-line search is usually based on keywords linked by the standard Boolean Operators (AND, OR, NOT). These operators allow the user to identify records where a given set of terms co-occurred in designated fields, but they cannot show how the concepts are interrelated .The on-line search of UKMARC file with the search statement "Teacher AND Student AND Assessment " will retrieve the documents on two different subjects as stated above. This can be largely avoided if the user can call for the display of a pre-coordinate subject statement as a kind of filter, during search procedure.

Like PRECIS dates as a difference (coded with $d) are not used in all cases. Historical periods may be listed in the input string from the DDC schedule.

In Literature, for example, the periods subdivisions enumerated in the DDC are used so that the index entry and the class number coincide. The methods associates with the generation of COMPASS index entries are same as that of PRECIS index entries.

The COMPASS is based on two basic components:

1) Subject Authority File; and
2) Thesaurus containing the semantic network of related terms, which serve as a source for *'see'* and *'see also'* references.

1) **Subject Authority File**

Subject Authority File developed by using the Washington Library Network (WLN) software contains two types of records:

i) *Term Authority Records (TAR)*, which contain terms representing single concepts—topical (including forms) or geographic (as subject and also as forms). Here, term may be either unitary (single-word) or compound (multi-word) expression of simple subject concepts. As for examples,

Unitary term heading: *Drugs*
Compound term heading: *Drug abuse*

ii) *String Authority Records (SAR)*, which contain one type of string only (i.e. topical). Each string consists of combination of two or more terms expressing complex subject. As for example,

String heading: *Drug abuse. Prevention*

A subject string once assigned may be reused for indexing documents of the same subject. Terms taken from TAR are arranged alphabetically under the given term along with the DDC number. The subject index of BNB refers to a DDC number in the following manner:

Hinduism
Jainism *related to* Hinduism 200

The above index entry looks like a relative index, first devised by Melvil Dewey.

In the main part of BNB, the COMPASS string appears at the end of the entry for bibliographic record. The printed subject index of BNB drawn according to COMPASS appears to be much shorter and user-friendlier at the expense of the precision.

2) **Thesaurus**

The semantic aspects of COMPASS are governed by the PRECIS network of related terms (i.e. the RIN file), which served as the source of *"see"* and *"see also"* references in the printed subject index of BNB. Subject Authority File developed for COMPASS by using the software WLN is also based on the concept of open–ended vocabulary, which mean that a new term can be admitted into the indexing vocabulary as soon as it is encountered in documents. This file also incorporates entries from the RIN file as appropriate.

COMPASS and UKMARC

When a document is handled by the office responsible for descriptive cataloguing, its detail is recorded on a standard worksheet, organised by MARC fields. Not all the information in the MARC record is relevant to the subject of a document, but certain fields are related specifically to classification number and others are related to subject headings. The different fields of the worksheet containing subject information of BNB are as follows:

050 Library of Congress (=LC) classification number

082 Dewey Decimal Classification (=DDC) number

600 Personal name as subject

610 Corporate name as subject

650 Topical subject headings

651 Geographical subject headings

660 Subject topical descriptors

661 Subject geographical descriptors

Fields 690, 691 and 692 used earlier for PRECIS string, RIN and Subject Indicator Number (SIN) respectively were discontinued: and the fields 660 and 661 replaced, with the introduction of COMPASS. Proper names as subject are located in the UKMARC fields 600 and 610 and are found in the Author/Title index, marked "t" to distinguish them from the same name as author or title. With the introduction of COMPASS, BNB discontinued inclusion of Library of Congress Subject Headings (LCSH), which was constructed from 6XX fields in the MARC records. LCSH contained headings with subdivisions, and headings subdivided according to "pattern headings" — neither of which was in the printed version of LCSH.

COMPASS and DDC

The indexer who writes the COMPASS input string also assigns the appropriate DDC number in the 082 field. The initial step of subject analysis is done only once while preparing the COMPASS input string for a document and this input string is taken as the basis for all later decisions relating to the derivation of the

subject data for a given document, and their incorporation in the relevant fields in the worksheet meant for the BNB. The DDC numbers serve three different purposes:

a) DDC numbers organise entries for bibliographic records in the classified part of the BNB and thus ensure collection of entries.

b) DDC numbers are also used as a source of feature headings. Usually, up to three levels of headings from DDC number are given, but sometimes for four and even five levels of heading from the DDC number are given. There are no feature headings for more details subdivisions. The index string drawn according to the COMPASS appears at the end of the main body of the entry, in italics. This system of producing feature heading has been reported to be unsatisfactory from the user point of view. In this connection, it is to be pointed out here that prior to the introduction of COMPASS, feature headings were constructed by the computer from the terms in PRECIS string selected in input order.

c) DDC numbers are now linked directly to the bibliographic records rather than through the subject string. Prior to the introduction of COMPASS, the PRECIS strings were used to generate the DDC numbers and also the feature heading for the BNB classified sequence, so that users could find their way through the file without being unduly reliant on the class numbers — both these are incorporated in the SIN file. The COMPASS has not been used for this purpose at all.

COMPASS and BNB

As stated earlier, the COMPASS string appears at the end for bibliographic record of a document in the classified/main part of the BNB. The subject index of BNB refers to a class numbers in the following manner:

Library operations
Classification compared with indexing 025

In the classified part of the BNB, a number of entries or bibliographic records has been arranged under the class numbers 025. The above mentioned subject index directs the user to scan entries under the class number 025 in the subject/ main part of the BNB in order to find out the one which has at the end of the subject heading "Classification compared with indexing".

The printed subject index of BNB appears to be much shorter and more user-friendly than the earlier one. A C Foskett has observed that in the BNB of 1990 (using PRECIS index), there are ten "see also" references for a term "Antiquities" and these are followed by 54 subject index entries, each modified by terms either in the qualifier or in the display to give a coextensive subject statement, leading to 54 different classes. In the BNB of 1991 using COMPASS index, the heading "Antiquities" appears at 28 places in the classified sequence leading to 28 different class numbers. To help the user, the subject index gives the context in which a particular piece of notation represents a given term. This looks like a relative index first devised by Melvil Dewey. In effect, every entry in the index is a "see" reference, taking us from the concept in verbal plane to its notational plane—that is, heading used for arrangement of entries for bibliographic records in the classified or main part of the BNB.

Any system needs time for its testing and development. With the introduction of COMPASS, BNB stopped including LCSH headings until protests from customers

and cooperative partners which finally led to their reintroduction in 1995. At the end of 1996 the British Library's own COMPASS system of indexing was dropped from the BNB with little or no protest from anybody and LCSH was further adopted as its sole subject access system.

4.3.8 POPSI

In 1964, Ranganathan demonstrated a new line of thinking regarding verbal indexing based on facet analysis [Ranganathan, 1964]. Since then, continuous research in this new line of thinking was going on, and ultimately the researches of Dr G Bhattacharyya on this new line of thinking led to three distinct but interrelated contributions in the field of subject indexing in India, which together constitute what has been designated as the General Theory of Subject Indexing Language (GT-SIL). Dr Bhattacharyya first explained the fundamentals of subject indexing languages with an extensive theoretical background in the GT-SIL. Three interrelated contributions, which constitute the GT-SIL, are as follows:

1) A methodology for structural analysis of names-of-subject developed through logical abstraction of the structures of outstanding SILs—such as those of Cutter, Dewey, Kaiser and Ranganathan. This methodology was based on the Deep Structure of Subject of Indexing Languages (DS-SILs).
2) A methodology for designing a specific purpose-oriented SIL, known as POstulate-based Permuted Subject Indexing or POPSI. The POPSI was developed through logical interpretation of the postulates for structural analysis of the names-of-subject forming part of the GT-SIL.
3) A methodology for designing a vocabulary control device, known as classaurus—an elementary category-based faceted scheme of hierarchical classification in verbal plane having all the necessary attributes of a conventional alphabetical thesaurus. It was designed on the basis of the postulates governing the structural analysis of names-of-subject forming part of the GT-SIL.

Structures of Subject Propositions

For the purpose of designing a SIL, the following types of structure in a name-of-subject (subject–proposition) have been recognized by Dr. Bhattacharyya:

1) ***Semantic structure:*** It refers to the structure in the dimension of denotation or comprehension. This structure is intrinsic to subject-propositions and is based on "Genus- Species", "Whole-Parts", and "Interfacet" relationships.
2) ***Elementary structure:*** It is the structure in the dimension of elementary categories to which the different substantive constituents of subject-propositions belong. This structure is artificially postulated and is recognized on the basis of the semantic significance of the substantives.
3) ***Syntactic structure:*** It is the structure in the dimension of the horizontal sequence of the elementary constituents of subject-propositions meant to generate the intended pattern of grouping in their vertical sequence.

Deep Structure of SIL (DS-SILs)

According to the GT-SIL, the structure of a specific SIL may be deemed to be a surface structure of the deep structure of SIL. By logically abstracting the results of analyses of the surface structures of the different SILs, it has been possible to arrive at the deep structure of SIL. The DS-SILs may be presented schematically as shown in Figure 4.4.

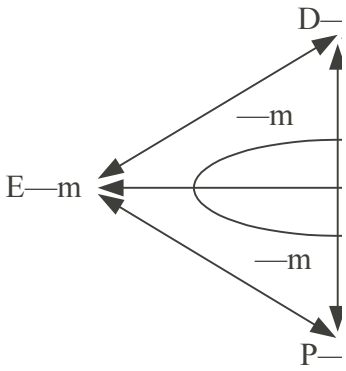


Fig. 4.4 : Deep structure of SIL

Annotation

It appears from the above figure that any specific substantive idea (and also the term denoting it) used to formulate a name-of-subject belongs to any one of the following elementary categories (D, E, A, P) and modifier:

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

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

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

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

5) *M=Modifier: The term 'm*odifier' refers to an idea used to qualify the manifestation of any one the elementary categories D, E, A and P, without disturbing its conceptual wholeness. As a result, it decreases the extension and increases the intension of the qualified manifestation without disturbing its conceptual wholeness. A modifier can modify a manifestation of any one of the elementary categories, as well as two or more elementary categories. Any manifestation of any elementary category may serve as the basis for deriving a Modifier. Modifiers generally create Species/Types. Modifiers are of two types:

   a) *Common Modifiers:* They refer to Space (e.g. Indian Music), Time (e.g. 15th Century Drama), Environment (e.g. Desert Birds), and Form (e.g. Dictionary of Physics). Common modifiers have the property of modifying a combination of two or more elementary categories.

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

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

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

Some other working concepts associated with the GT-SIL are explained below:

Organising Classification and Associative Classification

According to the GT-SIL, classification is a combination of both organising classification and associative classification. In other words, an indexing system is a combination of both organising classification and associative classification. The tasks involved in creating an organising classification are the categorisation of concepts and their organisation with the help of the DS-SILs. In organising classification compound subjects are based on genus-species, whole-part, and other inter-facet relationships. Here, classification is used to distinguish and rank each subject from all other subjects with reference to its COordinate—Superordinate—Subordinate—COllateral (COSSCO) relationships. It is achieved by means of interpolation and extrapolation of successive superordinates of each manifestation. The result of organizing classification is always a hierarchy.

In associative classification, a subject is distinguished from other subjects based on the reference of how it is associated with other subjects without reference to its COSSCO relationships. The result of associative classification is always a relative index.

Base and Core

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

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

Postulates Associated with the DS-SILs

The DS-SILs and its following associated postulates may be considered as the essence of the GT-SIL:

1) *Basic Sequence:* The sequence D followed by E (Either modified or unmodified) appropriately interpolated or extrapolated by A and P (either modified or unmodified) is a logical sequence of the elements of a Basic Chain manifesting in a compound subject-proposition. Any A or P may have A and/or P directly related to it. Their position is always after the A or P to which they are related. This sequence may be used as the basis for generating organising classification; and which in turn may form the basis of associative classifications. The logic behind this proposition is based on the purpose of generating a "General to Specific" sequence in the vertical arrangement of subject propositions.

2) *Source Organising Classification:* The basic sequence of manifestations augmented by the interpolation and extrapolation of successive superordinates of each manifestation, whenever warranted, gives rise to a Basic Modulated Chain; which can generate a source organising classification in alphabetical arrangement with the aid of suitable of apparatus introduce for this purpose. The basic modulated chain can be manipulated to generate chains meant for different organising classifications, and different associative classifications. This implies that there is no single absolute version of organising or associative classification. Classification is always purpose-oriented that determines the optimally efficient and effective versions of classification. The common organising structure of the manipulated versions of basic modulated chains constitutes the surface structure of specific SIL.

3) *Inter-facet relationship in Basic Chains:* In basic chains, the relationship between any two manifestations is the inter-facet relationship. For example, in the basic chain *"Agri-culture, Rice, Harvesting",* the relationship between any two manifestations is inter-facet relationship.

4) *Modifyee-Modified Relationship:* In basic modulated chains, the relationship between the modifyee and the modified manifestation is the Genus-Species relationship. For example, the relationship between "Disease" and "Infectious disease" is of this type.

5) *Associative Classification Effect:* The simple cycle permutation of each of the sought terms, in any style, with the indication of structure of subject propositions meant for organising classification has every effect of an associative classification.

6) *Systematic Grouping:* Only the notational representations of modulated chains can ensure in arrangement the ideal systematic grouping by juxtaposition. Only the alphabetical arrangement of modulated chains with suitable notational apparatus is the closest approximation to the purely notational grouping. Grouping by referencing has the organising classification effect.

7) *Synonyms, quasi-synonyms, and antonyms* can be controlled only by referencing, or by some acceptable substitute for referencing.

Characteristic Features of POPSI-Basic

The DS-SILs and its associated postulates are to be used as the basis for designing a Basic SIL, and it has been called POPSI-Basic. The postulates, principles and the working concepts of POPSI are the same as those forming part of the General Theory of SIL. POPSI as a process or operation for preparing subject-propositions consists primarily of (a) Analysis; (b) Synthesis; and (c) Permutation. The work of analysis and synthesis is primarily based on the postulates about the DS-SILs, and the work of permutation is based on the cyclic permutation of each term-of approach, either individually or in association with other terms for generating associative classification effect in alphabetical arrangement. The task of analysis and synthesis for POPSI–Basic is largely guided by the following POPSI-table. This table prescribes the notations that can be used to indicate mostly the categories of manifestation as well as the position of manifestations in subject-propositions. This feature of POPSI table is essential for mechanizing the horizontal arrangement of manifestation in subject-propositions, as well as for the vertical alphabetisation leading to organising classification.

Table 4.2: POPSI Table

0	Form Modifier		
1	General Treatment		
2	Phase relation		
2.1	General		
2.2	Bias		
2.3	Comparison		
2.4	Similarity		
2.5	Difference		
2.6	Application		
2.7	Influence		
	Common Modifiers		
3	Time modifier		
4	Environment modifier		
5	Place modifier		
6	Entity (E)		
		. 1 Action (A)	, Part
			. Species/Type
7	Discipline (D)	. 2 Property (P)	— Special modifier
Preceded by the notation of the manifestation in relation to which number is to be it is related		*Note:* A & P can go with another A&P also. In that case, the Action/Property. Preceded by that of its manifestation.	
7	Core (C)		
8	Base (B)	Features analogous to 6 Entity / 7 Discipline / Action / Property.	

Principles of Sequence for POPSI-Basic

The logical sequence of manifestations prescribed for subject-propositions meant for POPSI-Basic is the Basic sequence as prescribed in the postulates associated with the DS-SILs. A Species (Types)/Part follows immediately the manifestation in relation to which it is a Species (Type)/Part. A modifier (m) follows immediately the manifestation in relation to which it is a Modifier (m). The notation to be assigned to the manifestation of Discipline (D) in POPSI-Basic is always omitted .The ordinal value of this omitted notation is always taken to be the highest. The same procedure is adopted in case of any Base (B), whether it is a manifestation of D or not.

In subject-propositions for organizing classification, the different manifestations are normally arranged horizontally in the decreasing sequence of the ordinal values of the notations indicating their respective categories.

When arranged horizontally in the decreasing sequence of their ordinal values, the punctuation marks hyphen (-), full stop (.), and comma (,) fall in the same sequence. The horizontal sequence of the manifestations Modifier, Species/Type, and Part in a subject-proposition for organizing classification is the same as that mentioned above. In the vertical sequence of subject-propositions their ordinal values fall in the following increasing sequence: comma (,), full-stop (.), and hyphen(-).

Steps in POPSI-Basic

1) Analysis of the subject-proposition (=Analysis): It consists of identifying the elementary categories and modifiers in the statement of the subject.

2) Formalisation of the Subject-Proposition (=Formulation): It consists of the formalisation of the subject-proposition on the basis of the results of the step 1 (Analysis) according to the principles of sequence of components indicating the status of each facet.

3) Standardisation of the Subject-proposition (=Standardisation): It consists of deciding the standard term especially for the manifestation having synonyms.

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

5) Preparation of the Entry for Organising Classification (=Preparation of EOC): It consists of preparing the entry for generating organising classification by juxtaposition of entries in alphabetical sequence. [*Note*: The operation in this step is guided by the above mentioned POPSI-table].

6) Decision about Terms-of –Approach (=Decision about TA): It consists of deciding the terms-of-approach for generating associative classification for its effects; and of controlling synonyms.

7) Preparation of Entries of Associative Classification (=Preparation of EAC): It consists of preparation of entries under each term-of-approach, their individually, or in association with other terms, by cyclic permutation of sought terms for generating associative classification effect in alphabetical arrangement in such a way that under each term, a more or less systematic arrangement of subject-proposition is found.

8) Alphabetical Arrangement of Entries (=Alphabetisation): It consists of arranging all the entries in alphabetical sequence.

Demonstration of the Procedure for the POPSI-Basic

0 *Subject Indicative Expression*

Implications of technologies changes in rice cultivation for India

1) *Analysis*

D=Agriculture

E=Rice

A to E =Cultivation

P of A to E =Technology

P of P of A to E=Changes

P of P of P of A to E =Implications

m of the whole subject proposition=India

2) *Formalisation*

Agriculture (D), Rice (E), Cultivation (A to E), Technology (P of A to E), Changes (P of P of A to E), Implications (P of P of P of A to E), India (m of the whole subject proposition).

3) *Standardisation*

Agriculture (D), Field crop. Rice (E), Cultivation (A to E), Technology (P of A to E), Change (P of P of A to E), Implication (P of P of P of A to E), India (m of the whole subject proposition).

4) *Modulation*

Agriculture (D), Field Crop. Rice (E), Cultivation (A to E), Technology (P of A to E), Change (P of P of A to E), Implication (P of P of P of A to E), Asia, India (m of the whole subject proposition).

5) *Preparation of EOC*

Agriculture 6 Field Crop. Rice 6.1 Cultivation 6.1.2 Technology 6.1.2.2 Change 6.1.2.2.2 Implication 5 (for) Asia, India

6) *Decision about EAC*

Use each term other than 'Agriculture', 'Change', and 'Implication' as term-of –approach.

7) *Preparation of EAC*

Field Crop

Agriculture 6 Field Crop. Rice 6.1Cultivation 6.1.2 Technology 6.1.2.2 Change 6.1.2.2.2 Implication 5 (for) Asia, India

[*Note:* Similarly, EAC is to be prepared under each of the other terms-of–approach. Alternatively, making it multifaceted by using terms from upper links especially when warranted can make the heading of an entry for associative classification more specific. For example, the heading with 'Technology' as the term-of-approach may, in this case, consists of the following: "Technology, Cultivation, Rice"].

8) *Alphabetisation*

Arrange the entries according to the alphabetical order by terms-of-approach.

POPSI-Specific

As the situation prevails today it has become imperative for each and every information centre including a library to design and develop its own SIL in order to satisfy the condition of achieving its own specific objective. There has always been a tradition to depend upon a designer of a SIL. It is always found to be inadequate to satisfy the need at the local level. No standard SIL has yet been found to be suitable and adequate to meet the specific requirements of depth indexing in an area of micro subject. The basic assumption leading to the development of POPSI-Specific is that subject indexing is always a specific requirement-oriented activity. Difference in requirement would call for difference in syntax of the subject proposition; flexibility should be its rule, not the rigidity. Based on this assumption, POPSI tries to find out what is logically basic, and amenable to systematic manipulation to meet specific requirement. The POPSI-Basic is a product of the application of the GT-SIL and it is readily amenable to the systematic manipulation to generate purpose-oriented oriented specific versions known as POPSI-Specific. POPSI-Specific is always a derivation from the POPSI-Basic according to special decisions and rules to meet specific requirements at the local level. It may be noted here that this approach is totally different from that of earlier contributors of different SILs.

Principles of Sequence for POPSI-Specific

The postulates, principles and working concepts of POPSI–Specific are the same as those of POPSI–Basic forming the part of the GT-SIL. When the Base (B) and Core (C) do not correspond to those of POPSI–Basic, the guidance about the sequence of manifestations is readily available from the POPSI–Basic itself. The general principle in this case would be as follows: B (either modified or unmodified), is followed by C (either modified or unmodified), and the rest of the manifestations in the sequence as suggested in the postulates of basic sequence associated with the DS-SILs forming part of the GT-SIL.

Demonstration of the Procedure for POPSI-Specific

Let us consider the earlier example. Suppose that the purpose is to bring all or major portion of information pertaining to 'Cultivation Technology' together irrespective of its application to specific crops. In such a situation, 'Cultivation', which is a manifestation of Action (A), and "Technology", which is a manifestation of Property to Action (P to A), according to POPSI-Basic, are to be fused together, and the resulting manifestation 'Cultivation Technology' is to be treated as the Base (B). Suppose further that within the Base 'Cultivation Technology' we intend to organize information by 'Soil Preparation Technology', 'Sowing Technology', 'Weeding Technology', 'Harvesting Technology', etc. In that case they are to be treated as Sub-Base of the Base 'Cultivation Technology'.

Suppose still further that we intend to organise all these technologies by specific crops. In that case, the sub-bases are to be sharpened further by using each type of crop and specific crop as a Modifier of the Base and the Sub-Bases. When these are the primary decisions, the resulting indexing is called 'POPSI-Specific'; and it is implemented by systematic manipulation of POPSI-Basic. It will be evident from the following demonstration:

0 *Subject Indicative Expression*

Implication of technological changes in rice cultivation for India.

1 *Analysis*

Base (B) = Cultivation Technology

Modifier of the Base (m of B) = Rice

Property (P) of the (extended) B (P to B) = Changes

P of P of B = Implication

m of the whole subject proposition = India

2 *Formalisation*

Cultivation Technology—Rice (Extended B), Changes (P of B), Implication (P of P of B), India (m of the whole subject proportion).

3 *Standardisation*

Cultivation Technology—Rice (B), Change (P of B), Implication (P of P of B), India (m of the whole subject proposition).

4 *Modulation*

Cultivation technology—Field crop. Rice (B), Change (P of B), Implication (P of P of B), Asia. India (m of the whole subject proposition)

5 *Preparation of EOC*

Cultivation technology—Field crop. Rice. 9.2 Change 9.2.2 Implication 5 (for) Asia, India

6 *Decision of TA*

Use each term other than 'Cultivation technology', 'Change', and 'Implication' as a term of approach.

7 *Preparation of EAC*

Field Crop

Cultivation technology—Field crop. Rice 9.2 Change 9.2.2 Implication 5 (for) Asia, India

[Similarly under each of the other sought terms-of-approach, EAC is to be prepared].

Alternative multifaceted heading:

Field crop, Cultivation Technology

8 *Alphabetisation*

Arrange the entries according to the alphabetical order by terms-of-approach.

[Note: Evidently; this POPSI-Specific has been derived from POPSI-Basic by a systematic manipulation based on the following decisions:

a) Elimination of the discipline term "Agriculture"; and

b) Recognition of "Cultivation Technology" as the Base (B)].

Self Check Exercises

3) "Kaiser started from the point where Cutter left"—Discuss.

4) Mention the different types of relationships and their respective relational operators as proposed J.E.L. Farradane in his Relational Indexing.

5) How are the syntactical and semantic relationships dealt with in PRECIS?

6) Discuss the basic assumptions that led to the development of POPSI-Specific.

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

.....

.....

.....

.....

.....

.....

.....

4.4 POST-COORDINATE INDEXING SYSTEMS

Basically all the indexing methods are, by nature, coordinate indexing. The purpose of using a combination or coordination of component terms is to describe the contents of the documents more accurately. In case of pre-coordinate indexing system, coordination of component terms is carried out at the time of indexing (i.e. at input stage) in anticipation of the users' approach. In such a system, the most important aspect is to determine the order of significance or citation order by following the syntactical rules of the given indexing language. Rigidity of the significance order may not meet all the approaches of users. To satisfy the varieties of approaches of all the users, there is the provision for multiple entries

by rotating or cycling of the component terms representing the subject of the document. It has been observed that even the acceptance of multiple entry system covers only a fraction of the possible number of the total permutation, which in turn, results into the failure of the index file to provide a particular pattern of combination which the user is looking for. Consequently, a large portion of probable approach points is left uncovered. In pre-coordinate indexing system, both the indexer and the searcher are required to understand the mechanism of the system—the indexer for arriving at the most preferred citation order and the searcher for formulating an appropriate search strategy—in order to achieve the highest possible degree of matching of concepts. Pre-coordinate indexing system is non-manipulative. The searcher has no choice but to follow the citation order specified by the indexer. The above considerations and difficulties stemming from the pre-coordination of terms led to the development of post-coordinate indexing system, sometimes referred to as Coordinate Indexing System.

In post-coordinate indexing system, component concepts (denoted by the terms) of a subject are kept separately uncoordinated by the indexer, and the user does the coordination of concepts at the time of searching (i.e. at the output stage). The searcher has wide options for the free manipulation of the classes at the time of searching in order to achieve whatever logical operations are required. As the coordination of concepts are done at the output (search) stage this is known as Post Coordinate Indexing System. The Post-coordinate indexing system is of two types:

Term Entry System: Here, index entries for a document are made under each of the appropriate subject headings and these entries are filed according to alphabetical order. Under this system, the number of index entries for a document is dependent on the number of component terms associated with the thought content of the document. Here, terms are posted on the item. Searching of two files (Term Profile and Document Profile) is required in this system. For examples, Uniterm, Peek-a-boo, etc.

Item Entry System: It takes the opposite approach to term entry system and prepares a single entry for each document (item), using a physical form, which permits access to the entry from all appropriate headings. Here, items are posted on the term. Item entry system involves the searching of one file (i.e. Term Profile) only. For example, Edge-notched card.

4.4.1 Uniterm

Martimer Taube devised the Uniterm indexing system in 1953 to organise a collection of documents at the Armed Services Technical Information Agency (ASTIA) of Atomic Energy Commission, Washington. The system is based on concept coordination, where each component term (uniterm) is independent of all other terms and serves as a unique autonomous access point to all relevant items in the collection. Uniterm indexing system had a number of distinctive characteristics:

- Indexing by single words only;
- Terms are extracted from the text of the document indexed;
- No control over those terms;
- Indexing, being reduced to word extraction, can be conducted by relatively low-level personnel.

However, immediately after its development it was realized that concepts can be represented always by a single term. As a consequence it can be herself switched over to Unit concepts rather than Unit form. So sometimes this is also referred to as Unit Concept Indexing.

Here each document is assigned an accession number or 'address'. The thought content or subject of the document is analysed into uniterms. For each document an abstract is prepared and is recorded on the card along with the selected uniterms. The accession number assigned to a document is also written on the card. The card also contains bibliographical details of the document. These cards are arranged according to the ascending order of the accession numbers of the documents and are called *Numerical File.*

For each uniterm, a term card is prepared and it is divided into 10 equal vertical columns (from 0 to 9) in which accession number of the document relevant to the uniterm are recorded according to the system of terminal digit posting. Terminal digit of the accession number determines the column of its posting. Term cards are arranged according to the alphabetical order. This is called the *Uniterm File* of the system.

When a search is made, the subject of the search is broken into uniterms and the pertinent uniterm cards are pulled out from the alphabetical deck of the *Uniterm File.* Uniterm cards thus pulled out are matched to find the common accession number(s). The number(s) common in all such uniterm cards represent the sum total of the component concept of the specific subject. With the help of the common accession number(s), relevant card(s) are pulled out from the *Numerical File* where full bibliographical information of the required document(s) is available.

The merit of the uniterm indexing system is its simplicity and the ease with which persons without much knowledge of subject indexing can handle it. The limitations of the system are two-fold: (a) it takes much time in searching a document because of the searching of two files—uniterm file and numerical file particularly when it is practiced manually. Due to development of technology, it is done faster now; (b) it may retrieve irrelevant documents due to false coordination of uniterms. For example, searching with the uniterms 'science' and 'philosophy' may retrieve the documents on both the subjects 'Philosophy of Science' and 'Science of Philosophy', one of which might be irrelevant to a particular user. To overcome the problem, Engineering Joint Council devised `Roles' which may be attached to the terms to resolve the ambiguity.

4.4.2 Optical Coincidence Card / Peek-a-boo

Peek-a-boo is the trade name of the optical coincidence card. It is also called 'Batten Cards'. One card may punch 500 to 10000 accession numbers depending on the size of the card and size of the ruled squares. Instead of putting accession number (as is done in uniterm system), an item is indexed by punching a hole into the appropriate position that serves to represent the document number. The following tools or equipments are required for indexing under this system:

- For punching: Hand drill (i.e. hand punch) or electrically operated model;
- For lighting: Light box—A box containing different sources of light.

An optical coincidence card is shown below:

TERM

Fig. 4.5 : A sample of Optical Coincidence Card

The scanning is simplified with a beam of light that passes through the holes in the relevant cards (which are hold before the light) indicating accession number. Here one has to detect that through which hole(s) the light is passing in order to ascertain the accession number.

Advantages

1) Uniterm relies on our ability to notice the matching accession number(s) on the card(s) we are scanning having the probability of mistake. This can be avoided here.
2) They have the potential to store a relatively large number of documents.
3) It permits speedy search.

Disadvantages

1) There are chances of punching mistakes.
2) Withdrawal of card is very difficult because of the holes. Although colour cards are inserted for this purpose.
3) Like uniterm indexing system, it involves two level searches—first term profile and then documents profile.

4.4.3 Edge-Notched Card

Indexing on Edge-Notched card is based on punched card system. Their value is limited to very small collection.

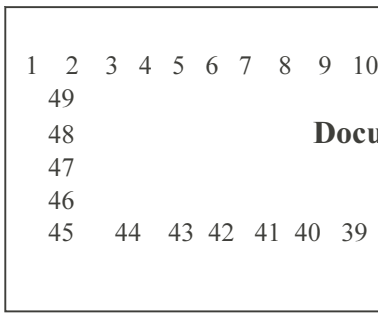
Features

1) One card corresponds to one document.
2) Number of holes on a card varies from 75 to 128.
3) A list of terms (term profile) associated with the thought content of the incoming documents is prepared along with their serial numbers.
4) Each hole represents a particular index term.
5) Subject index of a document is made by a series of notches punched out from the holes along the margin of the card.
6) The middle point of the card is used for bibliographic description of the document.
7) Equipment (physical apparatus) requires:
   - Wedge-shaped punch for notching marginal holes;
   - Needle(s) as sorting device (i.e. punched and non-punched cards)
8) Searching is carried out by inserting needle through the hole(s) corresponding to the request representation carries out searching and thereby allowing the cards punched in that position to fall.

Example: Let us consider a document on *'Chemotherapy of heart cancer among children'*. The term profile representing the serial numbers of the component terms associated with the thought content of this document appears to be as follows:

Serial No.	Terms
7	Cancer
15	Chemotherapy
28	Children

For this, one will have to punch the holes corresponding to the positions 7, 15 and 28, as shown below:



In zatacoding system of Calvin Mooers, relevant cards are sorted out by vibration after the selector needle is inserted into the pack.

Advantages

We have seen in the preceeding sections that two-level search is required in case of term entry system—i.e. searching term profile first and it is to be followed by searching the document profile. Edge-Notched Card System avoids second search.

Disadvantages

One card may contain up to a certain number of holes and as a result this system is not suitable for large collection.

These methods (4.4.2 & 4.4.3) are semi-mechanical methods and are no longer practiced.

4.4.4 Post-Coordinate Searching Devices

There are a number of devices, which are an essential part of post-coordinate search strategy to avoid the false drops, i.e. retrieval of unwanted documents because of the false coordination of terms at the time of searching. Devices used for the elimination false drops are used both for indexing and searching. They are discussed below:

a) **Use of Bound Terms**

Two or more terms in a subject may be bound to get rid of false coordination.

b) **Roles**

The role indicator is a symbol attached to the index term at the indexing stage to indicate the sense and use of the term in a particular context. Compilers of retrieval languages may develop their own roles. It may be shown that a term is functioning as 'Raw material', 'Product / Output', or that it is functioning as an 'Agent / Tool'. The role indicator is most useful in avoiding recall of terms with incorrect function. For instance, roles produced by the Engineering Joint Council (EJC) may be added to *'Alcohol'* to distinguish their functions as raw material, agent or tool, and product or output. The examples given below illustrate the use of EJC role indicators with the term *'alcohol'* in a practical situation:

Role	***Document***
1 [Raw Material / Input]	Use of *alcohol* in the production of wine
2 [Product / Output]	Manufacturing of *alcohol*
3 [Agent / Tool]	Use of *alcohol* as reagent in laboratories

c) Links

Links are special symbols used to group all the related concepts in a document separately, so that inappropriate combinations of terms are not retrieved. For example, a document number 327 on *'Welding of copper pipes and heat treatment of steel structures'* would be indexed:

WELDING	327A
COPPER	327A
PIPES	327A
HEAT TREATMENT	327B
STEEL	327B
STRUCTURES	327B

where the letters A and B indicate which terms are associated, thus ensuring that the document is not retrieved in a search for 'Welding of steel'.

d) Weighting

In weighting system, values are allocated to indexing terms or search terms according to their importance in particular documents or search programs. Weighting acts as a precision device and also as an output-ranking device. It is the method of the application of quantitative figures to the indexing / searching terms depending upon their emphasis in the document. The following table shows the weight worked out by Maron and others:

Table 4.3: Weightage System (by Maron)

Weight	Description	When Used
8/8	Major Subject	The term is highly specific and covers an entire major subject of the document.
7/8	Major Subject	The term is specific and covers most of a major subject of the document.
6/8	More Generic Subject	The term is too broad and covers a major subject.
5/8	Other Important Terms	Terms that would be used in a binary indexing system but not a major subject.
4/8	Less Generic Subject	The term relates to, but is too narrow to cover a major subject.
3/8	Minor Subject	Includes such terms as relate to results of experiments, intermediate methods, possible uses, etc.
2/8	Other Subjects	Other relevant tags.
1/8	Barely Relevant	Subject classifier would not want to use, but feels that some users might consider relevant.

The examples given below illustrate the use of weights from the above table:

Weight	*Term*	*Document*
8/8	X-Ray	Introduction to *X-Ray*
3/8	X-Ray	*X-Ray* treatments of the heart diseases

The second Cranfield Project used a less complex weighting system:

9/10 For concepts in the main theme of the document

7/8 For concepts in a major subsidiary theme

5/6 For concepts in a minor subsidiary theme

Weighting of search terms may also be a guide when adjusting search strategies. For instance, when reducing the coordination level, terms with lower weights may be excluded first, in the expectation that this will give optimum recall improvement.

Self Check Exercises

7) Distinguish between pre-coordinate and post-coordinate indexing systems?

8) What are the devices used to eliminate false drops in post-coordinate indexing?

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

.....

.....

.....

.....

.....

.....

4.5 AUTOMATIC INDEXING

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

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

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

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

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

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

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

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

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

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

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

4.5.1 Manual Indexing vs. Computerised Indexing

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

(Table contd.)

Manual Indexing	Computerised Indexing
5) Scanning, analyzing the critical views, understanding the concepts and using indexer's own subject knowledge and previous experience do indexing.	5) Computer cannot do this. It involves less intellectual effort.
6) Selected index terms less in number.	6) Selected index terms are more in number.
7) It is time consuming.	7) It takes less time.
8) It is expensive.	8) Index entries can be produced at lower cost.
9) It is very difficult to maintain consistency in indexing.	9) Consistency in indexing is maintained.

4.5.2 Methods of Computerised Indexing

4.5.2.1 Keyword Indexing

An indexing system without controlling the vocabulary may be referred as 'Natural Language Indexing' or sometimes as 'Free Text Indexing'. Keyword indexing is also known as Natural Language or Free Text Indexing. 'Keyword' means catch word or significant word or subject denoting word taken mainly from the titles and / or sometimes from abstract or text of the document for the purpose of indexing. Thus keyword indexing is based on the natural language of the documents to generate index entries and no controlled vocabulary is required for this indexing system. Keyword indexing is not new. It existed in the nineteenth century, when it was referred to as a 'catchword indexing'. Computers began to be used to aid information retrieval system in the 1950s. The Central Intelligence Agency (CIA) of USA is said to be the first organization to use the machine-produced keywords index from Title since 1952. H P Luhn and his associates produced and distributed copies of machine produced permuted title indexes in the International Conference of Scientific Information held at Washington in 1958, which he named it as Keyword-In-Context (KWIC) index and reported the method of generation of KWIC index in a paper. American Chemical Society established the value of KWIC after its adoption in 1961 for its publication 'Chemical Titles':

KWIC (Keyword-In-Context) Index

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

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

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

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

The operational stages of KWIC indexing consist of the following:

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

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

c) KWIC routine serves to rotate the title to make it accessible from each significant term. In view of this, manipulate the title or title like phrase in such a way that each keyword serves as the approach term and comes in the beginning (or in the middle) by rotation followed by rest of the title;

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

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

f) Arrange the entries alphabetically by keywords.

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

Control of damages of rice by insets 118

Damages of rice by insets / Control of 118

Insets / Control of damages of rice by 118

Rice by insets / Control of damages of 118

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

Variations of KWIC

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

KWOC (key-word out-of-context) Index

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

Control	Control of damages of rice by insets	118
Damages	Control of damages of rice by insets	118
Insets	Control of damages of rice by insets	118
Rice	Control of damages of rice by insets	118

Sometime, keyword is printed as heading and the title is printed in the next line instead of the same line as shown above. For examples,

Control
Control of damages of rice by insets 118

Damages
Control of damages of rice by insets 118

Insets
Control of damages of rice by insets 118

Rice
Control of damages of rice by insets 118

KWAC (key-word Augmented-in-context) Index

KWAC also stands for 'key-word-and-context'. In many cases, title cannot always represent the thought content of the document co-extensively. KWIC and KWOC could not solve the problem of the retrieval of irrelevant document. In order to solve the problem of false drops, KWAC provides the enrichment of the keywords of the title with additional keywords taken either from the abstract or from the original text of the document and are inserted into the title or added at the end

to give further index entries. KWAC is also called enriched KWIC or KWOC. CBAC (Chemical Biological Activities) of BIOSIS uses KWAC index where title is enriched by another title like phrase formulated by the indexer.

Other Versions

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

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

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

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

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

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

Uses of Keyword Index

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

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

Advantages

1) The principal merit of keyword indexing is the speed with which it can be produced;

2) The production of keyword index does not involve trained indexing staff. What is required is an expressive title coextensive to the specific subject of the document;

3) Involves minimum intellectual effort;

4) Vocabulary control need not be used; and

5) Satisfies the current approaches of users.

Disadvantages

1) Most of the terms used in science and technology are standardized, but the situation is different in case of Humanities and Social Sciences. Since no controlled vocabulary is used, keyword indexing appears to be unsatisfactory for the subjects of Humanities and Social Sciences;

2) Related topics are scattered. The efficiency of keyword indexing is invariably the question of reliability of expressive title of document as most such indexes are based on titles. If the title is not representative the system will become ineffective, particularly in Humanities and Social Science subjects;

3) Search of a topic may have to be done under several keywords;

4) Search time is high;

5) Searchers very often lead to high recall and low precision; and

6) Fails to meet the exhaustive approach for a large collection.

Self Check Exercise

9) What are the different versions of keyword indexing?

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

.....

.....

.....

.....

.....

4.5.2.2 Other Methods of Automatic Indexing

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

a) **Statistical Analysis**

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

a) *Weighting by location*

A word appearing in the title might be assigned a greater weight than a word appearing in the body of the work.

b) *Relative frequency weighting*

This is based upon the relation between the number of times the words is used in the document being indexed and the number of times the same word appears in the sample of other documents.

c) *Use of noun phrase*

Noun and adjective noun phrases can be selected as index terms and these are selected from the title or abstract of the document.

d) *Use of thesaurus*

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

e) *Use of association factor*

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

f) *Maximum-depth indexing*

This procedure indexes a document by all of its content words and weights these words, if desired, by the number of occurrences in the document. In this way, the problem of selecting term is avoided.

b) **Syntactic and Semantic Analysis**

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

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

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

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

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

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

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

Advantages

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

Disadvantages

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

4.5.3 File Organisation

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

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

File organisation forms an important element in computerised indexing. File organization is a technique for physically arranging the records of a file on secondary storage devices. This is the technique for organisation of the data of a file into records, blocks and access structures.

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

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

A schematic view of file organisation techniques is represented in figure 4.6.

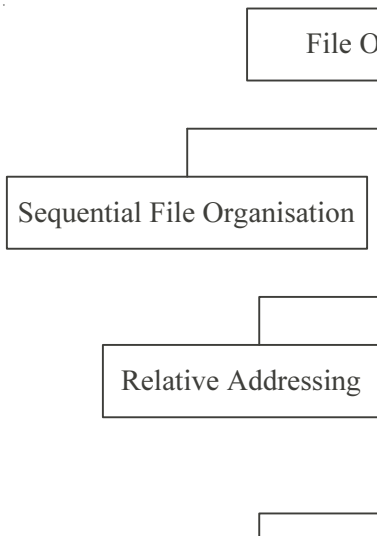


Fig. 4.6: Schematic view of file organisation

1) *Sequential File Organisation:* In this technique, records are stored in some predetermined sequence, one after another. One field, referred to as the primary key, usually determines their sequence of order.

2) *Direct File Organisation:* This technique supports direct access (also called random access), in which records can be accessed instantaneously and in any order from the data scattered throughout the disk.

   2.1 *Relative Addressing:* It is the simplest method of finding a record. Here, a record's primary key is associated with a specific physical storage location and contents of the records are stored in this address.

   2.2 *Hashing:* It is a method for determining the physical location of a record. Here, the record key is processed mathematically, and another number is computed that represents the location where the record will be stored.

   2.3 *Indexing:* It is a procedure for locating a record in a file stored randomly throughout the disk. Here, a primary index associates a primary key with the physical location in which a record is stored.

      2.3.1 *Ordered Index:* It is based on a sorted ordering of values.

      2.3.1.1 *Primary Index:* The records in the indexed file can be stored in some sorted order. If the file containing the records is sequentially ordered, the index whose search key specifies the sequential order of the file is called the primary index.

      2.3.1.2 *Secondary Index:* Indexes whose search key specifies an order that is different from the sequential order of the file are called secondary indexes.

      2.3.1.3 *B-Tree Indexes:* It takes the form of a balanced tree in which every path from the root to the tree leaf is of the same length. It eliminates the redundant storage of search key values.

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

Further details about File Organisation is dealt within Unit 2 of MLII-104.

4.5.4 Indexing Systems using AI Techniques

Artificial Intelligence (AI) is a branch of computer science concerned with study and creation of computer systems that exhibits some form of intelligence, systems that can learn new concepts and tasks, systems that reason and draw useful conclusions about the world around us, systems that can understand a natural language or perceive or comprehend a visual scene and systems that perform other types of feats that require human types of intelligence. The research in AI applications to information retrieval is gaining much importance. The recent development of the AI applications in the areas of which only those concerned with the subject indexing system is discussed below:

4.5.4.1 NLP-Based Subject Indexing System

Natural Language of Processing (NLP), one of the areas of AI, has gained momentum in research activity that explores how natural language text that is entered into a computer system can be manipulated and transformed into a form more suitable for further processing for improved storage and retrieval of information. NLP offers a potential viable alternative to statistical techniques in computerized indexing. Most automatic retrieval systems based on NLP techniques, convert the contents of the document files and user's queries in an internal form and the task of matching takes place at that level of the system, which is known as natural language interfaces, or *front-end systems.* At the core of any NLP technique there lays an important issue of natural language understanding. The process of building computer programs that understand natural language requires knowledge of how the words are formed, how the words in turn form clauses and sentences. In general, the different levels of knowledge that is to be used in natural language understanding falls into the following groups:

1) *Morphological Knowledge:* This level gives knowledge of word formation and deals with the morphological structure of words like the word root, prefix, suffix and infixes. The basic unit in a written word is a morpheme.
2) *Lexical Knowledge:* A lexicon consists of words considered as valid in the given domain. The lexicon may also contain syntactic markers or certain categories, which can be useful in processing. This level deals with thesaurus look up, spelling variations, acronyms and abbreviations, etc.
3) *Syntactic Knowledge:* Syntax deals with the structural properties and validity of input sentences—how a right combination of words in a particular sequence constitutes a valid sentence.
4) *Semantic Knowledge:* Semantics deals with the meaning of words and that of sentences. Different methods of representation of meaning have been developed over the years.
5) *Pragmatic Knowledge:* A given concept may occur in a number of different meanings, and to decide about the correct meaning in a given context the NLP system needs pragmatic knowledge. Pragmatic level deals with sentences in a particular context. This requires a higher-level knowledge, which relates to the uses of sentences in different contexts. This knowledge is useful because it helps to eliminate ambiguities and supplements the semantic representation.
6) *World Knowledge:* In order to carry out effective communication, both the communicator and communicate should have background knowledge, either to send or to receive a message, without any noise. This background knowledge is considered as the world knowledge of a particular domain.

Chomsky's classification of the types of grammar formalism may be correlated

with the above mentioned levels of knowledge. A NLP system having pragmatic and world knowledge constitutes his *'type 0'* grammars; systems with semantic knowledge or context sensitive grammars constitute his *'type 1'* grammars; systems with syntactic knowledge or *context free grammars* constitute *'type 2'* grammars; and systems with most restrictive or regular grammars constitute 'type 3' grammars. One method to formalize our knowledge is to provide a series of rewrite rules (known as grammars) that will generate the legal sentences of the language. Of the above mentioned grammars, *context free grammars* or *'type 2'* grammars, introduced by Noam Chomsky, are well understood from the computational point of view. In particular, the *Phrase structure grammars* (*PSGs*) have been used extensively in developing parsers. Another variant of it, called *Definitive Clause Grammars* (*DCGs*) is the basis for a programming language. *PROLOG (Programming in Logic)* and *LISP (List Processing)* are most popular languages in artificial intelligence programming.

For syntax analysis in NLP, valid sentences of a language are recognised and their underlying structures are determined. A central component of the syntactic phase is the *parser* (the term *parse* is derived from the Latin phrase *pars orationis* which means 'part of speech), a computational process that takes individual sentences or connected texts and converts them to some representational structure useful for further processing. *Parsing* refers to the use of syntax to determine the functions of the words in the input sentences in order to create a data structure that can be used to get at the meaning of the sentence. The major objective of the *parsing* is to transform the potentially ambiguous input phrase into an unambiguous form as an internal representation. A few of the notable computational models of the parsers are *Finite State Transition Networks (FSTN), Recursive Transition Networks (RTN), Augmented Transition Network (ATN), Definite Clause Grammar (DCG),* etc. There are many computational models of semantic grammars like *Conceptual Dependency Grammars (CDG), Modular Logic Grammars (MLG), Lexical Functional Grammars (LFG), Case Grammars (CG),* etc.

NLP has progressed to the point at which assignment indexing by computers should not be possible. Vleduts-Stokolov, for example, described an experiment performed at BIOSIS in which terms from a limited vocabulary of 600 biological concept headings were assigned automatically to journal articles. The assignment is achieved by matching article titles against a semantic vocabulary containing about 15,000 biological terms, which, in turn, were mapped to the concept headings. In NLP-based subject indexing, phrases are automatically extracted from texts to serve as content indicators. Nonsyntactic methods are based on simple text characteristics, such as word frequency and word proximity, while syntactic methods selectively extract phrases from parse trees generated by an automatic syntactic analyser.

It seems very likely that we will see increased emphasis on the use of natural language in information retrieval in future. This claim seems justified due to the following factors:

a) The continued growth in the availability of machine-readable databases many of which are in natural language form;

b) The continued expansion of online system, which is likely, eventually, to put terminal in the offices and houses of scientists and other professionals. Bibliographic searching in one of many possible applications of these terminals and natural language mode of searching seems imperative in this type of application;

c) A number of evaluation studies have indicated that natural language may offer several advantages over controlled vocabularies in many retrieval systems;

d) Natural language indexing systems have been shown to work, and to work well, in the legal field, the scientific information dissemination centres, the defence and intelligence communities; and

e) New development in computer storage devices will make the storage of very large text files increasingly feasible.

The development of the AI techniques in other two important areas concerned with the subject indexing system includes the following:

4.5.4.2 Knowledge Representation-based Subject Indexing System

Knowledge representation connotes a way of codifying knowledge. The arrangement and representation of knowledge as reflected in classification schemes and thesauri may be treated as one form of codification of knowledge. The codification helps computer to store, process and inference the codified knowledge. A computer system capable of understanding a message in natural language would seem to require both contextual knowledge and the processes for the inferences assumed by the generator. Some progress has been made towards computer systems of this sort, for understanding spoken and written fragments of language. Fundamental to the development of such systems are certain ideas about structures of contextual knowledge representation and certain techniques for making inferences from that knowledge. Based on the idea that knowledge is symbolic, the methods of using NLP tools and techniques include use of symbols and symbol structures, semantics and roles, categories and relations present in the subject statement. The work in the area of 'semantic primitives' or 'semantic factoring' leads to several means of representing knowledge for reasoning, including symbolic equations, frames and semantic nets.

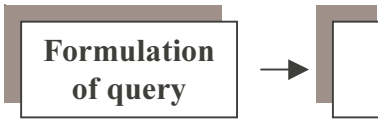
4.5.4.3 Expert System-based Subject Indexing

An *Expert system* is the embodiment, within the computer, of a knowledge-based component derived from an expert in such a form that the computer can offer intelligent advice or take an intelligent decision about the processing function. An *expert system* consists of knowledge acquisition, knowledge base system, inference machine, and user interface. This is one of the major areas of AI with a wide application to information processing and retrieval. The researches in this area lead to the development of expert systems, that should be fruitful in solving the problems of indeterminacy in both indexing and term selection for retrieval. Expert systems were designed to assist the user in query formulation and selection of relevant documents. A number of studies were undertaken in devising expert systems to aid the process of subject indexing. One notable example is the Medical Indexing Expert System or MedIndEx System (previously referred to as the Indexing Aid System) being developed at National Library of Medicine (NLM), USA. Referred to as a 'prototype interactive knowledge-based system', it uses an experimental frame language and is designed to interact with trained MEDLINE indexers by prompting them to enter MeSH terms as 'slot fillers' in completing document-specific indexing frames derived from the knowledge-base frame. Another expert system for machine-aided indexing used in the Central Abstracting and Indexing Service (CAIS) of the American Petroleum Institute (API) incorporates a rule-based expert system founded on the API thesaurus, which has been in operation since 1985. Terms automatically generated by the system (by matching abstracts against a knowledge-base derived from the API thesaurus) and by API's human indexers are reviewed by a human editor, and edited terms are added to print indexes and the online index. At present the base contains 14,000 text rules.

4.5.5 User Interface Design

User interface performs two major tasks – search or browse an information

collection and display of search results. It is also designed to perform other related tasks such as sorting, saving and/or printing of search results and modification of search query. Obviously, the success of any information retrieval system largely depends on the efficient and effective design of user interface. A well-designed user interface enhances the quality of interactions with information system and information search process my be divided into four major phases –



Shneiderman proposes following guiding principles for the design and development of user interface for computerised information retrieval –

- Strive for consistency in terminology, layout, instructions, fonts and colour
- Provide shortcuts for skilled users
- Provide appropriate and informative feedback about the sources and entity that is being searched for
- Design of message or notification system to indicate end of search process
- Permit reversal of actions so that users can undo or modify actions
- Allow users to monitor the progress of a search
- Permit users to specify the parameters to control a search
- Help users to recall their search queries
- Provide extensive online help and error-handling facilities
- Error messages should be clear and specific
- Provide facilities to enter long queries and use of Boolean, relational and positional operators
- Provide alternative interface for novice and expert users
- Provide multilingual interface if required

Information retrieval systems vary in terms of design, objectives, characteristics, contents, and users. However, the above guidelines with the help of following technical features will help to design an effective user-centred information retrieval system :

1) The presentation layer must reside within the client. Windows clients must contain an applications layer.
2) The system proposed must store all help files on the client for immediate retrieval/customisation.
3) The system proposed must enable the operator to request and receive context-sensitive help for the command in use with a single keystroke.
4) The client(s) proposed must enable the operator to page backwards and forwards through the help text, and include hypertext links to related topics, screen shots, examples, etc., and the ability to access an index of help topics.
5) The clients proposed must enable the operator to transfer any data field from one screen to another.
6) Wizards must guide the operator through a series of steps necessary to complete a defined process, without the use of commands or traditional menus.

7) Windows clients must perform all edit checking of user input before sending input to the server.

8) The clients proposed must display a form with all appropriate fields on the screen when an operator initiates a command.

9) The clients proposed must enable the user to type data onto the screen at the current cursor position.

10) The clients proposed must enable the user to use delete and insert character keys to correct mistakes.

11) The clients proposed must retain work forms on the screen, until the operator changes to another command.

12) If an error is detected, the system must report the error on the screen, leaving the form and the operator's input otherwise intact.

13) The system proposed must display an explanatory error message whenever the operator has provided inappropriate input.

14) The system proposed must set up a session with a user, which records user's search history and allows users to re-execute their previous searches against the same server or a different server or servers.

15) The system proposed must allow restriction of access to local or remote databases based upon the user's login and password.

16) The system must retain a user's authorization as he or she moves from one database to another within one session.

17) The system must allow both access by specific IP without sign-on and sign-on capability for users from outside the allowed IP range.

18) The proposed Web client must support creation and execution of simple or complex searches.

19) The proposed Web client must support browsing (SCAN) and selecting terms from standard vocabulary device.

20) The proposed Web client must support hypertext searching for related items and should support display of cross-reference information.

21) The proposed Web client must support sorting of search results to user-defined sequence.

22) The proposed Web client must support record export to printer, local file or e-mail.

23) The proposed Web client must support linking to library holdings information on multiple servers via the Z39.50 protocol.

24) The proposed Web client must support execution of previous searches from stored search history.

25) The proposed Web client must allow authentication of users by user ID and optional personal identification number, and permit authenticated users to:

    a) access their own user accounts to view the status of charges, holds, fines, and bills associated with their account; and

    b) obtain access to additional databases, such as licensed journal citation and other reference databases closed to anonymous users.

26) The system proposed must enable an authorized user to:

    a) save records to the review file one at a time, or in stated ranges;

    a) view records stored in a review file using Vendor's client software;

b) remove records from the review file;

c) duplicate records in the review file; and

d) print records in the review file;

An example of User Interface in computerised indexing is given below:

Fig. 4.7 : Screen snap shot of user interface

Self Check Exercises

10) Mention the methods adopted in measuring the word significance in computerized indexing system.

11) What are the different levels of knowledge used in natural language understanding for the NLP-based subject indexing system?

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

.....

.....

.....

.....

.....

.....

.....

4.6 NON-CONVENTIONAL INDEXING : CITATION INDEXING

Citation indexing, a variety of non-conventional indexing developed by Eugene Garfield, is the technique of bringing together the documents, which manifest association of ideas to establish the relevancy of information in a document

through mechanical sorting of citations in a citation index. A citation index, according to Garfield, is *an ordered list of cited articles each of which is accompanied by a list of citing articles.* The citing article is identified as a source, the cited article as a reference. When a document refers to some other documents it can be assumed that there is some similarity or association of ideas between the citing documents and cited documents. An author usually mentions along with his work all those documents, which he has consulted or referred to. This is called citing or referring documents. Citations are thus the references made to other documents in the text of a work. A document to which a reference is made is called a cited document, while a document, which makes reference to the cited document, is called a citing document. In general, a citation implies a relationship between a part and whole of a cited paper and a part or whole of the citing paper. The relationship between the cited documents and the citing documents forms the basis of citation indexing. By following the subsequent citations, the history of an idea can be traced—where and how it has been applied and whether it is sustained, rejected or absorbed into the latter work representing knowledge.

Though the citation indexing has provided a new approach to the bibliographic file organisation, the basic idea is not new. The first application of this idea was evident in *Shepard's Citations,* developed by Frank Shepard in 1873 as an index to American legal cases. It represents a list of American court cases, each case being followed by a complete history written in a simple code. The listing under each case displays the publications that have been referred to the cases, other court decisions that have affected the cases, and any other references that may be of value. Appreciating the usefulness of the system, Eugene Garfield stressed the need of adopting it in the field of science and technology in 1950's and in 1961 brought out an experimental Science Citation Index (SCI) after some preliminary studies. Institute for Scientific Information (ISI), Philadelphia in 1963, published the first SCI comprising the literature of 1961 and since then it is being brought out on regular basis. The online version of the SCI, known as SCISEARCH, is being published from 1974 and is available through several of the major online host systems—such as, DIALOG, etc. The compact disc version of the SCI is also available. It incorporates new feature called *related records*, which leads the searcher to other records having references in common with the ones he has already retrieved. ISI also brought out the Social Science Citation Index (SSCI) and Arts and Humanities Citation Index (A&HCI) in 1973 and 1978 respectively. The publication of the citation classics, with the first issue of *Current Contents* in 1977, forms an important and interesting venture of the ISI. Citation classics are articles selected every week by ISI from SCI, SSCI and A&HCI databases that have become classics in different fields by virtue of their high citation rate.

Science Citation Index (SCI)

SCI is the brain-child of Dr. Eugene Garfield, one time Director of the Institute for Scientific Information (ISI), Philadelphia. He created and nurtured it and brought it to the present position of the popularity, and is perhaps the best example of its kind. Initially, SCI started with two parts—*Citation Index* and *Source Index.* Since 1967, its third part, the *Permuterm Subject Index* was introduced. These three indexes, though separate are related to each other. It is published quarterly and the December issue is a cumulative issue for the whole year.

Citation Index (CI)

A CI entry contains two types of information— (a) information about a cited item (reference), and (b) information about citing items. It is arranged alphabetically by cited author, using last name of the first author (when there are more than one author). An entry for a cited item consists of the first author's

name and initials, year of publication of the cited item, details about the source document like title in abbreviated form, its volume number and starting page number. When different papers of a particular author are cited, the papers are arranged in chronological order. When different authors in different places cite a particular document, all these citing items are displayed immediately under the cited item by alphabetical order of the source authors. The information included about a citing item are: citing author's name and initials, title of the source document, its year of publication, volume number and starting page of the citing item. Coded symbols like A (for abstract), E (for editorial), N (for technical notes), etc. are used to indicate the nature of citing items.

The CI part also has two other sections on anonymous cited documents and *Patent Citation Index (PCI)*. Anonymous cited documents are arranged alphabetically by titles, and the PCI presents a list of all cited patents (foreign and domestic) in numerical sequence by patent number and usually provides the year of issuance, the inventor's name, and country.

Example (cited reference followed by citing references):

KOHARA Y				
87	CELL	50	495	
RAMAYO	GENETICS	122	65	89
WOOD	ER J BIOL CHEM	264	8297	89

Source Index (SI)

The SI is an author index of the citing items, arranged alphabetically by the last name of the first author of the source item. An entry for a citing item contains the name of the first author, the names of co-authors (up to ten), if any, the full title of the article, the title of the source document, its volume number, issue/part/ supplement number, starting page number, year, coded symbol for type of document, number of references in the bibliography of the citing item, accession number of the source periodical as filed at ISI, and author's address. The SI also has two other sections on anonymous items, and a *Corporate Index.* Anonymous items are arranged alphabetically by the titles of periodicals, and are given at the beginning of the SI. In the *Corporate Index,* all the source items are arranged alphabetically by author under the name of the organization where the work was done. When more than one organization is involved in a particular project, an entry is created for each organization.

Example:

WOOD ER

MATSON SW – THE MOLECULAR-CLONNING OF THE GENE ENCODING THE ESCHERICHIA-COLI 75-KDA HELICASE AND THE DETERMINATION OF ITS NUCLEOTIDE-SEQUENCE AND GENETIC-MAP POSITION U 5453

J BIOL CHEM 264(14) : 8297-8303 89 54R

UNIV N CAROLINA, DEPT BIOL, CHAPEL HILL, NC 27599, USA

Permuterm Subject Index (PSI)

The PSI satisfies the conventional subject approach of the users. In the PSI, significant or subject denoting terms are selected from the titles of papers appearing in the source periodicals. These terms are then permuted so that each of these terms serves as a primary term. Co-terms—that is, all other terms, which are related to a primary term, are arranged alphabetically under the given primary term like relative index of DDC.

Example (primary term followed by co-terms):

MOLECULAR-CLONNING		
GENE	.....	COONEY A
	.....	HOFFMAN E
	.....	
GENETIC-MAP	.....	WOOD ER

Search Strategy

The basic technique implies that a searcher starts with a reference or an author, he has identified through a footnote, book, encyclopedia article, subject index, or through his personal knowledge. Then, a search is to be carried out in the CI part of the SCI to identify the authors citing this reference. A second search through the SI will reveal complete bibliographical descriptions of these source items, which will enable the searcher to judge the relevancy of the documents to his query. This search may be extended in order to build a more extensive bibliography for a particular inquiry through the process, what is called 'cycling'. Once the searcher locates the citing references in the SI, he can have a look on other documents of the citing author, which may be relevant to his query. Here we see how a search involving both CI and SI operates.

When a searcher is not aware of any reference, he may start search through Permutation Subject Index. For searching a PSI, a searcher is required to compile a list of terms relevant to the search topic. Then, he should have a look on the PSI to identify one of these terms used as primary term. Simultaneously he can locate the author using the term in his title. He can then enter the SI to identify complete bibliographical description of the document. Considering co-terms along with primary term in the PSI can sharpen this search. Once this more defined subject is obtained, he locates the appropriate authors to be used in the SI. Information thus obtained can be used in the CI to locate subsequent citing sources.

Advantages

1) This index enables to link up all recent papers on a subject with papers published earlier but with which the recent ones have association of ideas.
2) While the conventional subject indexes have problems relating to terminology, this index is unaffected by problems.
3) It has already been established that this index is faster and more efficient than conventional indexes for searching.
4) It provides the best access to the interdisciplinary literature.
5) The genesis of an interdisciplinary or newly emerging subject can be traced from this index.
6) With the help of this index an author of a cited document can know how far his ideas or research results have been appreciated, applied or criticized.
7) It can be used as a tool to identify the interconnected papers on a subject to trace the history and development of it.
8) It can also be used as a measuring tool for the progress of science, as a means of automatic content analysis and as a way of evaluating journals, papers, authors, departments, and candidates for awarding prizes, appointment and promotion.
9) It is a self-updating index as each new citation of an earlier paper is automatically listed making the list of citing documents up-to-date.
10) The preparation of the citation index needs no intellectual activity since its preparation is fully computerised.

Disadvantages

1) Citation indexes stand on the foundation of the citation practice of authors. Any discrepancy at this level will generate noise at the output stage.
2) It does not provide any logical or conventional subject arrangement to which the users are accustomed. A title not in harmony with the subject content is bound to affect the citation indexing.
3) It is often complained that a user for starting the search in a citation index must know at least one reference (PSI can be used to locate a starting point).
4) A citation index retrieves only the related documents and not the contents of the documents.
5) A search in the SCI retrieves high proportion irrelevant items from the point of view of a particular searcher.

Social Science Citation Index (SSCI)

ISI brought out SSCI in 1973. It is multidisciplinary index to the literature published in the world's leading 1400 journals in social sciences, and to selected relevant items from 3,100 science journals. SSCI covers almost all the social science subjects like Anthropology, Business, Communication, Criminology, Education, Geography, History, Information Science, Library Science, Law, Linguistics, Philosophy, Sociology, etc. It appears 3 times a year and has a calendar year index. The online version of the SSCI, known as SOCIAL SCISEARCH enables speedy and accurate searches of social science literature. The compact disc version of the SSCI is available individually for the year 1986 to 1991 along with a 5-year cumulation covering 1981—1985.

The structure, format and search of the SSCI are the same as SCI. With four integrated indexes—Citation Index, Source Index, Permuterm Subject Index, and Corporate Index, one can get access to vast amount of not only the social science literature, but also to the science literature having links with the social sciences.

Arts & Humanities Citation Index (A&HCI)

ISI brought out the A & HCI in 1978 after intensive two-year marketing research. A & HCI covers more than 25 arts and humanities disciplines like Archeology, Language, Architecture, Literature, Arts, Music, Classics, Philosophy, Dance, Religious studies, Films, Television & Radio, Folklore, Theatre, History, Theology, etc. It is published three times a year.

The structure, format and search of the A & HCI are the same as those of SCI and SSCI. Like SCI and SSCI, A & HCI consists of Citation Index, Source Index, Permuterm Subject Index, and Corporate Index.

Self-Check Exercise

12) What are the different parts of Science Citation Index?

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

.....

.....

.....

.....

.....

4.7 WEB INDEXING

The Web (also known as the World Wide Web or W3) is a vast collection of files accessible to the public through the Internet, viewed through a browser, and connected by hypertext links. The information available on the Web is so vast and easily accessible; the library professionals cannot afford to ignore it. The Web consists of a very large number of things called documents, identified by names called Uniform Resource Locators (URLs). As on today the WWW contains more than three billion pages of publicly accessible information. Anyone can publish anything on the Internet. When we search the Web we can find endless examples of information that is inaccurate, inaccessible, invalid, incomplete, out of date, unreadable or simply irrelevant. Searching the Web is compared to dragging a net across the surface of an ocean. The Web was developed to link related material from different information sources. The current use of explicit links implies that the Web is static. On the contrary, the Web is highly dynamic with new information constantly becoming available (and 'old' information being removed). While an author can include links to related documents that exist at the time of writing, new documents may become available in the future. The reasons for increasing difficulties in finding the required information on the Web in time may be summed as :

- The number of data is very large.
- The mass of data is increasing rapidly.
- The data always change in structure and content.

4.7.1 Meaning and Scope

The term 'Web indexing' refers to (a) search engine indexing of the Web, (b) creation of metadata, (c) organisation of Web links by category, and (d) creation of a Website index that looks and functions like a back-of-book index. It will usually be alphabetically organised, give detailed access to information, and contain index entries with subdivisions and cross-references. In the most general sense, Web indexing means providing access points for online information materials which are available through the use of World Wide Web browsing Software. The key to this is the use of human intellectual input to analyse and categories materials, rather than reliance on computerized searching tools. It appears from the Proceedings of the Web Indexing Workshop held at the University of Melbourne during July 2000, the following issues fall within the purview of web indexing:

a) Uploading of 'traditional' indexes (and the documents to which they refer) on to the Web to provide a wider audience with access to them.

b) 'Micro' indexing of a single Web page, in order to provide users with hyper linked access points to the materials on the page.

c) 'Midi' indexing of multiple pages, largely or wholly contained within a single Web site and falling under the responsibility of a single Webmaster.

d) 'Web-wide' indexing, providing users with centralized access to widely scattered materials, which fall under a single heading (e.g. every web page dealing authoritatively with 'breast cancer').

e) 'Macro' schemes designed to simplify or unify access to large number of Web pages falling under many different headings (e.g. every web page dealing authoritatively with any medical topic).

f) The addition of comments and annotations to provide users with some guidance before they link to selected sites and pages.

The absence of librarians' skills is well reflected in the indexing approach taken by the general search engines like Alta Vista, Google, Yahoo etc. which are at the most successful in recall but definitely not in precision. In addition, the output of these search engines definitely lacks uniformity. The search tools used in the Web usually include search engines, subject directories and information or subject gateways. To put the Web indexing in proper perspective it has become imperative to familiarise with how Web works?

4.7.2 Operational Aspects of the Web

a) A Web page or home page is a document of any size—from a few lines to an entire book. A Web site might have hundreds or thousands of pages. Every Web page has the URL (Uniform Resource Locator), which acts as the address of the file accessible on the Web. e.g. *http://www.vidyasagar.ac.in/DLIS/srr/index.htm/*. Here, *http* (Hypertext Transfer Protocol) is the protocol being used to access the resource; *www.vidyasagar.ac.in* is the name of the computer where the resource is located; *DLIS/srr/index.htm* is the pathname to locate the specific file, and this portion of the URL contains information needed by the protocol to specify which document is intended.

b) By using the *http* protocol, a program (most commonly a browser but any program can do this "\web agents" and spiders are examples of such programs that are not browsers) can retrieve a document whose URL starts with "*http//*:". The document is returned to the program, along with information about how it is encoded, for example, ASCII or Unicode text, HTML, images in GIF or JPG or MPEG or PNG or some other format, an Excel or Lotus spreadsheet, etc.

c) Documents encoded in HTML (HyperText Markup Language) form can contain a combination of text, images, formatting information, and links to other documents. Thus, when a browser (or other program) gets an HTML document, it can extract the links from it, yielding URLs for other documents in the Web. If these are in HTML format, then they too can be retrieved and will yield yet more links, and so on.

d) A *spider* is a program that starts with an initial set of URLs, retrieves the corresponding documents, adds the links from these documents to the set of URLs and keeps on going. Every time it retrieves a document, it does some (hopefully useful) work in addition to just ending the embedded links.

e) One particularly interesting kind of spider constructs an *index* of the documents it has seen. This index is similar to the index at the end of a book. It has certain key words and phrases, and for each entry it lists all of the URLs that contain that word or phrase. There are many kinds of indexes and the methods of deciding what words or phrases to index and how to extract them is at the cutting edge of research. Usually, every word in the document (except, perhaps, the very common words like "and", "the", "a", and "an") is indexed.

4.7.3 Pre-requisites for Web Indexing

i) *Know the Web*

It is necessary to be familiar with the dynamic and distributed nature of the information resources on the Web, the forces shaping it, and what makes an indexable resource.

ii) *Be a User*

A web indexer needs to be a user of the Web to look for good and bad examples of indexing and to keep track of the latest technological development.

iii) *Know How to Create the Content*

It requires knowledge about the HTML. One can look at the 'Source' of the Web page to see how it is done. The ability to mock up HTML pages will be of much help for the Web indexer to know the working of the HTML.

iv) *Know about Web Site Design*

What's good / what's bad, what works / what does not, a good index is useless unless it is compatible with the browser accessing it.

v) *Know about Search Engines* and how we can add value to them.

vi) *Start looking at XML* with particular reference to what does it mean for indexing.

4.7.4 Search Engines

Web search engines (SEs) came into existence in 1994. According to "Yahoo Search Engine Directory Listing, 2003", there are over 448 major search engines. Some examples of leading /search engines include:

- Alta Vista (www.altavista.com)
- Excite (www.excite.com/search)
- Google (www.google.com)
- Hotbot (www.hotbot.com)
- Northern Light (www.northernlight.com)
- Khoj (www.khoj.com)

A SE is a searchable database of Internet files collected by a computer program (called a wanderer, crawler, robot, worm, and spider). Indexing is created from the collected files, e.g. title, full text, size, URL, etc. There are no selection criteria for the collection of files. 'Search engines allow the user to enter keywords and the search engines retrieve Web documents from its database that match the keywords entered by the searcher. The search engine doesn't wait for someone to submit information about a site. Rather, it sends out robot programs or spider (Crawler, Web crawlers) that visits publicly accessible websites following all links it comes across collecting data for search engine 'indexes'. A spider discovers new sites and updates information from sites previously visited. A spider can also be used to check links within a website.

4.7.4.1 Components of Search Engine

A search engine might well be called a search *engine service* or a *search service.* As such, it consists of three components:

1) *Spider:* Programs that traverses the Web from link to link, identifying and reading pages.
2) *Index:* Web database containing a copy of each web page gathered by the spider.
3) *Search engine mechanism:* Software that enables users to query the index and that usually returns results in relevancy ranked order.

4.7.4.2 Types of Search Engines

A search engine downloads all the information that the page contains and then examines that information to index keywords and phrases that can be used to categorise the sites. The exact method that it uses to do this, and which information

it looks at to create the index, varies according to the search engine. Those words and phrases are added to the database alongside the URL and a description of the site. Search engines can be categorised into three types on the basis of the indexing techniques employed by them:

1) *Active Search Engine:* It collects web pages information by itself. It uses a program calls "Spider" or "Web robot" to index and categorise Web pages as well as Web sites. The spider travels around the WWW in search of new sites and adds entries to their catalogue.

2) *Passive Search Engine or Subject Directories:* Search engines of this type are possibly more accurately referred to as directories. It does not seek out web pages by itself. They rely on the WWW users to submit details on their site or their favourite sites in order to build up a database. Upon receiving the submissions somebody from the search engine company trots along to the Website suggested, has a look at it, and then places the details in the right part of the database by finding the main category and sub-categories into which website should fall. For example, Yahoo directory (www.yahoo.com) has 14 main subject categories and each of these categories has many sub-categories and those sub-categories also contain their own sub-categories, and so on almost ad infinitum. Hierarchically organised directories tend to be smaller than those of the search engines, which mean that result lists tend to be smaller as well. Because subject categories are arranged by category and because they usually return links to the top level of a Website rather than to individual pages, they lend themselves best to searching for information about a subject rather than for a specific piece of information.

   Due to the size of the Web and constant transformation, keeping up with important sites in all subject areas is humanly impossible. Therefore, a guide by a subject specialist to important resources in his area of expertise is more likely than a general subject directory to produce relevant information and is usually more comprehensive than a general guide. These guides are known as *Specialized Subject Directories***.** Such guides exist for virtually every topic. For examples:

   - Voice of Shuttle (http://vos.ucsb.edu) provides an excellent starting point for humanities research.
   - Film buffs should consider starting their search with the Internet Movie Database (http://us.imdb.com).

3) *Meta Search Engine:* An increasing number of search engines have led to the creation of "meta" search tools, often referred to as *multi-threaded search engines.* A meta search engine does not catalogue any web pages by itself. It simultaneously searches multiple search engines. When a query is put before this type of search engine, it forwards that query to other search engines. There are two types of meta search engines:

   a) One type searches a number of engines and does not collate the results. This means one must look through a separate list of results from each engine that was searched. It may present the same result more than once. Some engines require the searcher to visit each site to view the results, while others will fetch the results back to their own sites. When the results are brought back to the site, a certain limitation is placed on what is allowed to be retrieved. With this type of meta search engine, one can retrieve comprehensive, and sometimes overwhelming, results. An example of this type of engine is *Dogpile.*

   b) The other type is more common and returns a single list of results, often with the duplicate hits removed. This type of meta engine always brings the results back to its own site for viewing. In these cases, the engine retrieves a certain number of documents from the individual engines it has searched, cut off after a certain point as the search is processed. *Inference Find* claims to

return the maximum number of results that its targeted search engines will allow. Other meta search engines stop processing a query after a certain amount of time. Still others give the user a certain degree of control over the number of documents returned in a search. All these factors have two implications:

- These meta search engines return only a portion of the documents available to be retrieved from the individual engines they have searched; and
- Results retrieved by these engines can be highly relevant, since they are usually grabbing the first item from the relevancy-ranked list of hits returned by the individual search engines.

Some examples of meta search engines are:

a) Metacrawler (www.metacrawler.com)

b) SurfWax (www.surfwax.com)

c) Zapmeta (www.zapmeta.com)

Search engines can further be categorized by scope:

1) **General Search Engine:** It covers a range of services and facilitate Boolean search. Examples: Google, Alta Vista, etc.

2) **Regional Search Engine:** It refers to country specific search engine for locating varied resources region-wise. Examples: Euro Ferret (Europe), Excite uk (UK), etc.

3) **Subject Specific Search Engine:** It does not attempt to index the entire Web. Instead. it focuses on searching for Websites or pages within a defined subject area, geographical area or type of resource. Examples: Geo index (Geography / Environmental science), Biochemistry Easy Search Tool (Biochemistry). Because this specific search engine aims for depth of coverage within a single area, rather than breadth of coverage across subjects, they are often able to index documents that are not included even in the largest search engines databases. Some examples of subject specific search engines are:

www.123india.com	Regional
www.in.altavista.com	Regional
www.yahoo.co.uk	Regional
www.naukri.com	Employment
www.ndtv.com	News
www.zipcode.com	Weather
www.khoj.com	India-specific

4.7.4.3 Features of Search Engines

Search engines offer numerous features:

1) When using a Web search engine by entering more than one word, the space between the words has a logical meaning that directly affects the results of the search. This is known as the *default syntax.* For example: In Alta Vista, Infoseek, and Excite, a search on the words "birds migration" means that the searcher will get back documents that contain either the word "birds", the word "migration" or both. The space between the words defaults to the Boolean OR. This is probably not what the searcher wanted for this search. Again in HotBot, a search on the words "birds' migration" in Lycos and Northern Light means that the searcher will get back documents that contain both the words "birds" and "migration". Here, the space between the words defaults to the Boolean AND. Perhaps this is more appropriate.

2) Search engines return results in schematic order. Most search engines use various criteria to construct a term relevancy rating of each hit and present

the search results in this order. Criteria can include: search terms in the title, URL, first heading, HTML META tag; number of times search terms appear in the document; search terms appearing early in the document; search terms appearing close together; etc. Google page ranking algorithm displays mostly cited/hyperlinked Web sites/Web pages at the top of the screen.

3) One of the most interesting developments in search engine technology is the organization of search results by concept, site, domain, popularity and linking rather than by relevancy. Search engines that employ this alternative may be thought of as *second-generation search services.* For example:
   - Direct Hit ranks according to sites other searchers have chosen from their results to similar queries.
   - Google ranks by the number of links from pages ranked high by the service.
   - Inference Find ranks by concept and top-level domain.
   - MetaFind sorts results by keyword, alphabetically or by domain.
   - Northern Light sorts results into Custom Search Folders representing concepts and/or types of sites.
4) Often multiple pages are retrieved from a single site because all they contain the given search term. Alta Vista, Infoseek, HotBot, Northern Light and Lycos avoid this by a technique called *results grouping,* whereby all the terms from one site are clustered together into one result. It provides the opportunity to view all the retrieved pages from that chosen site. With these engines, one may get a smaller number of results from a search, but each result is coming from a different site.
5) Search engines do not index all the documents available on the Web. For example, most search engines cannot index files to password-protected sites, behind firewalls, or configured by the host server to be left alone. Still other Web pages may not be picked up if they are not linked to other pages, and are therefore missed by a search engine spider as it crawls from one page to the next. Search engines rarely contain the most recent documents posted to the Internet; do not look for yesterday's news on a search engine.
6) Contents of databases will generally not show up in a search engine result. A growing amount of valuable information on the Web is not generated from the databases. This aspect of the Web is sometimes referred to as "*the invisible Web*" because database content is "invisible" to search engine spiders.
7) Some search engines allow users to view a viewed display of the retrieved Web sites / Web pages, clustered under different topics related to the search term(s). Examples include Kartoo (http://www.kartoo.com), Vivisimo (http://www.vivisimo.com), etc.

4.7.4.4 Functions of Search Engines

There are differences in the ways various search engines work, but they all perform three basic tasks:

a) They search the Internet by using a specialized software, called crawler or robot; these software/agent can find out web pages by following hyperlinks;
b) These agent/software send the cached version of web pages to the repository of a search engine (SE) and SE keeps an index of the words they find, and where (URL) they find them; and
c) They allow users to look for words or combinations of words found in that index.

The above-mentioned functions of search engine may be illustrated through the diagram given below:

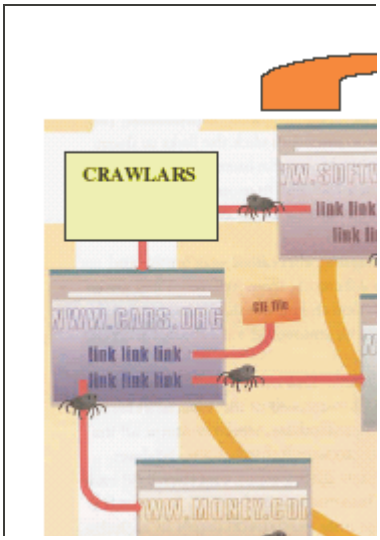


Fig. 4.8 : Diagramatic view of a search engine

4.7.4.5 Subject Directory Vs. Search Engine

We have already discussed in the section 4.7.4.2 of this Unit that a subject directory is a service that offers a collection of links to Internet resources submitted by the site creators or evaluators and organised into subject categories. Directory services use selection criteria for choosing links to include, though the selectivity varies among services. Most directories offer a search engine mechanism to query the service. When using subject directories, keep in mind that there are two basic types of subject directories: *academic and professional directories* often created and maintained by subject experts to support the needs of researchers, and *commercial portals* that cater to the general public and are competing for traffic. *INFOMINE* is a good example of an academic subject directory. *Yahoo* is a good example of a commercial portal. The points of differences between the SD and SE are shown in the following table:

Table 4.4 : Subject Directory Vs. Search Engine

Subject Directory (SD)	Search Engine (SE)
1) A SD provides categorized list of websites with brief description based on submission by Web site owners, scrutinized and edited by professional editors. It enables the searcher to move from menu to menu, making one selection after another until he gets to the level where the chosen sites are enlisted. Alta Vista, which is the largest SD today, covers more than 2 million Web sites.	1) While a SD categorizes Web sites and contains very littleinformation about them (just the description), a SE indexes all the information on all the web pages it finds. A SE deals with specific piece of information, not categories.
2) A SD takes the searcher to the home page of a Web site, and from there he can explore to get to what you want eventually.	2) A SE takes the searcher to the exact page on which thewords and phrases he is looking for.
3) With SD, a searcher requires to have a minimal knowledge to understand the catalogue.	3) With SE, one can start with one specific piece of information, like a name or phrase, and use that to find more, without knowing anything about the subject.
4) A SD is crafted by human beings, based on their judgment, like files in a file cabinet.	4) Indexes in the SE are generated automatically, based on the words and phrases that are found on Web pages. There is no human judgment filtering or rearranging of the information.
5) A SD is organised like a library.	5) Indexes in the SE are organized internally so that computer can help to find precise information from vast amount of knowledge.
6) A SD can be useful when a searcher has a vague idea of what he wants or a broad topic and wishes to view recommended sites relevant to that topic.	6) A SE is appropriate to use when a searcher is looking for a specific site or has narrow topic to pursue.
7) A SD categorises web sites and contains very little information about them (just description).	7) A SE indexes all the information on all the web pages it finds.
8) A SD requires having a minimal knowledge of the subject to understand the categories.	8)With SE, one can start with a specific piece of information, like a name or a phrase, and use that to find more, without knowing anything about the subject.

Because subject directories and search engines are so different and complement one another's capabilities, most major search engines have partnered with one or another of the major directories. For instance, Alta Vista now uses the *Open Directory Project* and the *LookSmart Directory* for its categories and Web sites. Today, the Open Directory includes over 1.5 million Web sites, in over 2, 50, 00 categories. Based on the work of over 23,000 editors, it is growing very quickly. For comparison, Yahoo now has about 1.2 million sites. Its growth is limited by what it is possible for a staff of paid editors to accomplish.

4.7.4.6 Subject Gateway

A subject gateway is recognised to be specialised in resources on a particular field and is compiled by the people, not robots. Subject Gateway or Information Gateway-type resources include Internet catalogues, subject directories, virtual libraries and gateways and these resources are organised into hierarchical subject categories. It put more focus on relevance and quality.

4.7.5 Semantic Web

The web has been said to reach its full potential only when it becomes a place where data can be shared, processed, and understood by automated systems as

well as by people. But the reality is that though the present day search engines like Google, Yahoo!, etc give us lot of hits, mostly spew out irrelevant information in answer to a search query. The problem with these search engines is that they use mostly statistical methods like frequency of occurrence of words, co-occurrence of words, etc. resulting a number of irrelevant hits against a search on the Web. Though some search engines like Google and Yahoo! use human edited entries; still they come up with a large number of wrong hits. The "semantic web" is an approach to extend the web with semantic information. Tim Berners-Lee, the developer of HTML, HTTP, URI and WWW, introduced the concept of Semantic Web. His visualisation of Semantic Web is that in future we will have intelligent software agents that will analyse a particular given situation and present us with the best possible alternatives. The idea behind Semantic Web is to develop such technologies that make the information more meaningful for the machines to process, which in turn makes search and retrieval of information more effective for searchers.

4.7.5.1 Meaning of Semantic Web

The dictionary meaning of the term "semantics" is "the study of meaning". It refers to the meaning of a string in some language, as opposed to syntax, which describes how symbols may be combined independent of their meaning. In the world of Internet the word "semantics" has slightly different meaning. In that world it means that both human and machine should have the understanding of particular topic. The World Wide Web Consortium (W3C) gives the following two definitions for the Semantic Web:

1) The Semantic Web is the representation of data on the World Wide Web. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs (Uniform Resource Identifiers) for naming.
2) The Semantic Web is the extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

The conception of Semantic Web is characterised by developing tools and technologies like languages, standards and protocols so that the Web becomes meaningful.

4.7.5.2 Key Technologies of Semantic Web

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

1) **Uniform Resource Identifier (URI):** A URI is an identifier used to identify objects in a space. The most popularly used URI is the URL. URL is a subset of URI, they are not synonymous. A URI can be anything from the name of a person or an email address or URL or any other literal. The URI specifies a generic syntax. It consists of a generic set of schemes that identify any document/ resource like URL, URN (Uniform Resource Name), URC (Uniform Resource Characteristic), etc.
2) **XML (eXtensible Markup Language):** The emergence of HTML led to the situation where nearly everybody wanted to publish his or her own HTML-based web pages on Internet. The content problem arose rather

quickly when it became clear that there is huge amount of information on the Net, but which is mostly unreachable because of the chaos of it: The HTML didn't tell anything of the nature of the data but only the proposed layout of it. So, anybody didn't really know anything about the structure of the data of web pages. XML was designed to attach semantic to data. The HTML is about displaying information while XML is about describing information. XML is a markup language for documents containing structured information. Structured information contains both content (words, pictures, etc.) and some indication of what role that content plays (for example, content in a section heading has a different meaning from content in a footnote, which means something different than content in a figure caption or content in a database table, etc.). XML is the first step in bringing meaning to the web. While it is similar to HTML, XML allows users to add arbitrary structure to their documents by defining their own tags, but says nothing about what the structures mean. Meaning is expressed by RDF, which encodes it in sets of triples, each triple being rather like the subject, verb and object of an elementary sentence. These triples can be written using XML tags. Many have misconception that XML will replace HTML but to whatever is the case finally actual representation is done in HTML format.

3) **RDF (Resource Description Framework)**: The problem of XML is that since it characterizes the data included into elements, it does so only by allowing the schema designer to select the names of the elements 'semantically'. The RDF is thought to be the next breakthrough technology since it provides a way to build an object model from which the actual data is referenced. RDF is a foundation of processing metadata. The most important feature of RDF is that it is developed to be domain-independent, i.e. it is very general in nature and does not restrict/apply any constraint on any one particular domain. RDF metadata expresses the meaning of term and concepts in a language (XML) that is understandable to computers. Actually the RDF relies on the support of XML. RDF is said to be a framework for describing and interchanging metadata. It has been proposed as a generic framework, which is applicable to many (or any) application domains.

RDF is suitable for enhance search engines, cataloguing for describing and classifying the content available in particular web site or digital library. It can be used by intelligent software agents to facilitate knowledge sharing and exchange and in content rating in classifying resources and describing collection of pages. RDF offers :

1) A standardised syntax for writing metadata and ontologies; and
2) A standardised set of modeling primitives like instance of and subclass of relationships.

In practice RDF is a set of triplets consisting of (i)Subject, (ii) Predicate, and (iii) Object.

In RDF things are described like: "A crane is a bird" (Crane is an instance of class Bird), when "a crane" is a subject, "is" is a predicate and "a bird" is an object. A document makes assertions that particular things (people, Web pages or whatever) have properties (such as "is a sister of," "is the author of") with certain values (another person, another Web page). Subject and object are each identified by a Universal Resource Identifier (URI). The verbs are also identified by URIs, which enables anyone to define a new concept, a new verb, just as defining a URI for it somewhere on the Web. The triples of RDF form webs of information about related things. Because RDF uses URIs to encode this information in a document, the URIs ensure that concepts are not just words in

a document but are tied to a unique definition that everyone can find on the Web. Data representation in RDF is built on the triplet or *Resource, Property* and *Value:*

- **Resource (Subject):** A resource is any entity that has to be described by a URL and is equivalent to *subject* in normal English grammar. It includes the entire web pages, as well as entire elements of an XML document. An example of a resource is a draft document and its URL is *http://www.textuality.com/RDF/Why.html*
- **Property (Predicate):** A Property is any characteristics of a Resource that has a name and is equivalent to *predicate* in normal English grammar. For example: a Web page can be recognised by 'Title' or a man can be recognised by his "Name'. So both are attributes for recognition of resource 'Web page' and 'person' respectively.
- **Value (Object):** A *property* must have a *value*, which is equivalent to *object* in normal English grammar.

The concepts related the resource, property and value of the RDF model can be described through the example of RDF based metadata encoding of web resources. The example below shows the description of the homepage of Vidyasagar University by using Dublin Core Metadata Element Set (DCMES) in the framework of RDF model. It applies DCMES in the description of the property set of the homepage of the said university.

Example of RDF Data Model

Resource	Property	Value
http://www.vidyasagar.ac.in	dc:title	Vidyasagar University Home page
	dc:creator	Chakraborty, B.
	dc:subject	Academic Institute
	dc:descriptipon	Home page for the Vidyasagar University, Midnapore, West Bengal. The University established in the year 1983 under UGC Act……..
	dc:publisher	Central Library, Vidyasagar University
	dc:contributor	Sarkar, A.K.
	role=content	
	wrtiter	
	dc:date	10/12/2001
	dc:format	text/html
	dc:identifier	http://www.vidyasagar.ac.in
	dc:language	En
	dc:coverage	Education and Research
	dc:rights	Vidyasagar University

4) **Ontologies**: XML and RDF deal with metadata i.e. they deal with the description of the information available on the Web. But, if machines are expected to interact with each other or share data in the true sense of the word, then semantic interoperability is essential. For this, a formal specification

is required to explicitly define various terms and their relationships. It is here that ontologies come into picture. An ontology, as used in the context of knowledge sharing, is defined as *a specification of a representational vocabulary for a shared domain of discourse—definitions of classes, relations, functions, and other objects—is called an ontology.* In simple terms, ontology can be said to be the definition of entities and their relationship with each other. It is based on semantic net used in Artificial Intelligence. Ontologies define data models in terms of classes, subclasses and properties. For instance, we can define a man to be a subclass of human, which in turn subclass of animals that is a biped i.e. walks on two legs.

Ontologies provide taxonomies and inference rules: The taxonomy defines classes of objects and relations among them. For example, an address may be defined as a type of location, and city codes may be defined to apply only to locations and so on. Ontology may express the rule "if a city code is associated with a state code, and an address uses that city code, then that address has the associated state code. Inference rules tell the computer how to deduce and derive knowledge by combining specific types of information about specific categories.

Ontologies can be categorized into two types: general ontologies (e.g. SENSUS, Cye, WordNet, etc.) and domain-specific ontologies (e.g. GALEN—Generalized Architecture for Languages, Encyclopedias, and Nomenclatures in Medicine; UMLS—Unified Medical Language System). Ontologies are built on using XML and RDF. But, during recent past, many specific Web ontology development languages have been developed. Some notable examples are: OIL (Ontology Interface Layer) and DAML (Distributed Agent Markup Language). These two languages have now merged into a single ontology language, called DAML+OIL. In late 2001 the W3C set up a working group called *WebOnt* to define an ontology language for the Web, based on DAML+OIL.

Ontologies can enhance the functioning of the Web in many ways. They can be used in a simple fashion to improve the accuracy of Web searches—the search program can look for only those pages that refer to a precise concept instead of all the ones using ambiguous keywords. More advanced applications will use ontologies to relate the information on a page to the associated knowledge structure and inference rules.

5) **Agents**: Another crucial element of the Semantic Web is the "agent" technology. Agents are software programs that contain logical algorithms that process information and make intelligent decisions based on a pre-programmed set of rules. Agents assist users in their tasks by getting / semantic information from ontologies and annotated web pages. Inference engines perform reasoning services for agents. There are already many examples of agents in use today. Search engines, for example, perform a set of logical processes based on the search criterion we supply them. It is hoped that many different agents will be able to work together to perform even more complex tasks on the Semantic Web than are performed by agents today.

The real power of the Semantic Web will be realized when people create many programs that collect Web content from diverse sources, process the information and exchange the results with other programs. The effectiveness of such software agents will increase exponentially as more machine-readable Web content and automated services (including other agents) become available. The Semantic Web promotes this synergy: even agents that were not expressly designed to work together can transfer data among themselves when the data come with semantics.

Self Check Exercises

13) Give the meaning and scope of 'Web indexing'.

14) Discuss the functions of the different components of a Search Engine.

15) What are the key technologies involved in the development of the Semantic Web?

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

.....

.....

.....

.....

.....

.....

.....

4.8 SUMMARY

This Unit provides a comprehensive view of theoretical and practical aspects of subject indexing. It is impossible to understand the subject indexing properly without being aware of the different principles and processes associated with it. For this, different aspects of indexing principles and processes are discussed at the outset. It is followed by the discussion on the significant contributions made in the areas of pre- and post-coordinate indexing systems at different points of time since the days of Cutter. Each contribution has been discussed with illustrative examples. The traditional subject indexing systems and techniques have taken a new turn with the applications of computers in 1950s. So, computerised indexing forms an important part of this Unit. The different methods and products of computerised indexing, its differences with manual indexing, user-interface design, indexing systems using Artificial Intelligence techniques like Natural Language Processing (NLP), Knowledge Representation Model and Expert System-based subject indexing systems are discussed here. This Unit also looks at the phenomenal growth of content on the Web as an indexing problem and describes the approaches currently used to index the Web resources. In this context, different search tools used in finding the resources on the Web along with the technologies developed so far to make computers understand the semantics underlying contents of the Web resources.

4.9 ANSWERS TO SELF CHECK EXERCISES

1) The processes of subject indexing consist of two stages: (1) establishing the concepts expressed in a document; and (2) translating these concepts into the components of the given indexing language. Establishing the concepts expressed in a document calls for understanding the overall content of the document by examining the important parts of the text like title, abstract, introduction, the opening phrases of chapters and paragraphs, illustrations, tables, diagrams and conclusions. /after examining the document, the indexable concepts are to be identified and then selected in the light of the purpose for which the indexing data will be used. Finally, it is to be followed by the translation of the selected concepts into the components of the given indexing language.

2) Exhaustivity is a measure of the extent to which all the distinct topics discussed

in a document are considered for indexing. In other words, it is a measure of the number of index terms per document in an indexing system. A high level of exhaustivity increases recall. Exhaustivity is a matter of indexing policy. Specificity, on the other hand, is the degree of preciseness of the subject to express the thought content of the document. Specificity is an intrinsic quality of the indexing language itself to represent the specific subject exactly and co-extensively. A high level of specificity increases precision.

3) C. A. Cutter was the first person that first gave a generalized set of rules for subject indexing in his *Rules for Dictionary Catalog,* published in 1876. Cutter provided rules for specific as well as compound subject headings. Cutter regarded subjects as specific and classes as broad. But, in practice, it was these broad classes that Cutter entered his specific subjects. Further, according to Cutter, order of the component terms in a compound subject heading should be the one that is decidedly more significant. But Cutter could not provide any guideline as to how one will come forward to decide which one is more significant. The question of significance varies from user to user and from indexer to indexer. J. O. Kaiser started from this point where Cutter failed to provide the guideline regarding the question of significance. Kaiser, in his *Systematic Indexing*, published in 1911, prescribed that the compound subject should be analysed by determining the relative significance of the different component terms of a compound subject through classificatory approach of categorization of terms. He categorized the component terms into two fundamental categories: *concrete* and *process.* Kaiser provided the guideline that *concrete* is more significant than *process* and so, he laid a rule that a *process* should follow *concrete.*

4) J. E. L. Farradane has identified the nine relationships in his *Relational Indexing.* These nine relationships and their respective relational operators are

   a) Concurrence / 0
   b) Self-activity / *
   c) Association / ;
   d) Equivalence / =
   e) Dimensional / +
   f) Appurtenance / (
   g) Distinctness /)
   h) Reaction / -
   i) Causation / :

5) Syntactical relationships in PRECIS are handled by means of a set of logical rules, role operators and codes. They regulate the organisation of terms in the input string by the indexer and their manipulation to generate index entries by the computer. Role operators act as instructions to the computer. Semantic relationships in PRECIS are regulated by a machine-held thesaurus that serves as a source of *See* and *See also* references in the index. A thesaurus is generated simultaneously with the preparation of input string.

6) The basic assumption leading to the development of POPSI-Specific is that subject indexing is always a specific purpose-oriented activity. But, there has always been a tradition to depend upon a designer of a subject indexing language, and such dependency always found to be inadequate to meet the specific requirement of subject indexing at the local level. Differences in requirement would call for differences in syntax of the subject proposition. It has been stated that the flexibility should be the rule of syntax, not the rigidity. Based on this assumption, POPSI tries to find out what is logically

basic, known as POPSI-Basic, and is readily amenable to the systematic manipulation to generate purpose-oriented specific versions, known as POPSI-Specific.

7) Pre-coordinate indexing involves coordination of component terms in a compound subject by the indexer at the time of indexing in anticipation of users' approach. In post-coordinate indexing, component terms in a compound subject are kept separately uncoordinated by the indexer, and the user does the coordination of terms in accordance with his requirements at the time of searching. In pre-coordinate indexing system, the rigidity of the significance order associated with the syntactical rules may not meet the approaches of all the users of the index file. But, in post-coordinate indexing, the searcher has wide options for the free manipulation of the terms at the time of searching in order to achieve whatever logical operations are required.

8) The following devices are used to eliminate false drops in post-coordinate indexing:
   a) Use of bound terms;
   b) Links;
   c) Roles; and
   d) Weighting.

9) Different versions of keyword indexing are:
   a) KWIC (Keyword –In-Context) Index;
   b) KWOC (Keyword Out-of-Context) Index;
   c) KWAC (Keyword Augmented-in-Context) Index;
   d) KWWC (Keyword-with-Context) Index;
   e) KEYTALPHA (Key Term Alphabetical) Index;
   f) WADEX (Word and Author Index);
   g) DKWTC (Double KWIC) Index;
   h) KLIC (Key-Letter-In-Context) Index;
   i) KWIT (Keyword-In-Title) Index; and
   j) SWIFT (Selected Words In Full Titles) Index.

10) Different methods adopted in measuring the word significance in computerized indexing are
   a) Weighting by location;
   b) Relative frequency weighting;
   c) Use of noun phrase;
   d) Use of thesaurus;
   e) Use of association factor; and
   f) Maximum-depth indexing.

11) Different levels of knowledge used in natural language understanding for the NLP-based subject indexing system falls into the following groups:
   a) Morphological knowledge;
   b) Lexical knowledge;
   c) Syntactic knowledge;
   d) Semantic knowledge;
   e) Pragmatic knowledge; and
   f) World knowledge.

12) There are three parts in Science Citation Index:
   a) Citation Index;
   b) Source Index; and
   c) Permuterm Subject Index.

13) The term 'Web indexing' refers to (a) search engine indexing of the Web,

(b) creation of metadata, (c) organization of Web links by category, and (d) creation of a Website index that looks and functions like a back-of-book index. It will usually be alphabetically organised, give detailed access to information, and contain index entries with subdivisions and cross-references. In the most general sense, Web indexing means providing access points for online information materials which are available through the use of World Wide Web browsing Software. The following issues also fall within the scope of Web indexing:

a) Uploading of 'traditional' indexes (and the documents to which they refer) on to the Web to provide a wider audience with access to them.

b) 'Micro' indexing of a single Web page, in order to provide users with hyper-linked access points to the materials on the page.

c) 'Midi' indexing of multiple pages, largely or wholly contained within a single Web site and falling under the responsibility of a single Webmaster.

d) 'Web-wide' indexing, providing users with centralised access to widely scattered materials, which fall under a single heading (e.g. every web page dealing authoritatively with 'breast cancer').

e) 'Macro' schemes designed to simplify or unify access to large number of Web pages falling under many different headings (e.g. every web page dealing authoritatively with any medical topic).

f) The addition of comments and annotations to provide users with some guidance before they link to selected sites and pages.

14) Functions of the different components of a search engine are :

a) **Spider**: Computer program that visit websites following all links it comes across collecting data for search engine indexes, identifying and reading Web pages;

b) **Index**: A searchable database containing indexing terms created from the Web pages by the spider; and

c) **Search engine mechanism**: Software that enables users to query the index and that usually returns results in relevancy ranked order.

15) Key technologies involved in the development of the Semantic Web are :

a) Uniform Resource Identifier (URI);

b) eXtensible Markup Language (XML);

c) Resource Description Framework (RDF);

d) Ontology; and

e) Agent software.

4.10 KEYWORDS

Amplified Phrase Order : The order of component terms in a phrase achieved by using the necessary prepositions in between them. It is a corollary of the order of significance provided by E J Coates.

Analet : Two or more isolates linked by relational operators according to J E L Farradane's Relational Indexing system constitute *Analet.*

Analysis : It refers to the conceptual analysis, which involves deciding what a document is about—that is, identification of different component ideas (concepts) associated with the thought content of the document.

Associative Classification: It refers to the association of a subject with other subjects without reference to its COSSCO relationships and results in a relative index.

Chain Indexing : A method of deriving alphabetical subject entries from the chain of successive subdivisions leading from the general to most specific level needed to be indexed. For this, it takes the class number of the document concerned from a preferred classification scheme for deriving subject index entries.

Citation Index : An ordered list of cited articles (references), each of which is accompanied by a list of citing articles (sources).

Citation Indexing : Techniques of bringing together the documents (cited documents) which manifest association of ideas to establish the relevancy of information in a document (citing document) through mechanical sorting of citations in a citation index. The relationship existing between the cited documents and citing documents forms the basis of *Citation indexing*.

Concrete : One of the fundamental categories propounded by Kaiser, which refers to things, place and abstract terms not signifying an action.

Consistency in Indexing : It is a measure that relates to the work of two or more indexers.

Controlled Vocabulary : A controlled vocabulary refers to an authority list of terms showing their interrelationships and indicating ways in which they may usefully be combined to represent specific subject of a document.

COSSCO Relationships : It refers to Coordinate-Superordinate-Subordinate-Collateral relationships in organising classification.

Default Syntax : A logical meaning of the use of space between the words when entering more than one word in a Web search engine carries out a search.

Exhaustivity : The use of enough terms to cover the all topics discussed in a document. It relates to the breath of coverage in indexing. It is sometimes called *depth indexing*.

Expert System : It is the embodiment, within the computer, of a knowledge-based component derived from an expert in such a form that the computer can offer intelligent advice or take an intelligent decision about the processing function.

eXtensible Markup Language (XML) : A subset of Standard Generalized Markup Language (SGML), a widely used international text-processing standard. XML is being designed to bring the power and flexibility of generic SGML to the Web, while maintaining interoperability with full SGML and HTML. XML is the first step in bringing meaning to the Web.

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

Hypertext Markup Language (HTML) : The standard text-formatting language for documents on the World Wide Web. HTML text files contain content that is rendered on a computer screen and markup, or tags that can be used to tell the computer how to format that content. HTML tags can also be used to encode metadata and to tell the computer how to respond to certain user actions, such as a mouse click.

Links : Special symbols used to group all the related concepts in a document separately for the elimination of false drops.

Index : A systematic guide to the contents of documents, comprising a series of entries, with headings arranged in alphabetical or other chosen order and with references to show where each item indexed is located

Indexing : The process of evaluating information entities and creating indexing terms, normally subject or topical terms, that aid in finding and accessing the entity. Index terms may be in natural language or controlled vocabulary or a classification notation.

Indexing Language : An indexing language is an artificial language consisting of a set of terms and devices for handling the relationship between them for providing index description. It is also referred to as a *retrieval language*.

Indexing Program : Computer software used to order things; frequently used to refer to software that alphabetizes some or all of the terms in one or more electronic documents.

Input String : It refers to a set of terms arranged according to the role operators in PRECIS.

Invisible Web : It refers to those pages what we can't see in the results pages after we run a search on the Web. Here, the pages remain hidden due to a variety of technical reasons, to search engines excluded them because the pages do not allow them access and also to pages, which restrict entry through the log-in password method.

Item Entry System : A type of post coordinate indexing system in which items are posted on the term. A single entry is prepared for each item, which permits access to the entry from all appropriate headings.

Keyword Indexing : An indexing system based on the usage of natural language terminology for deriving the index entries. Significant words denoting the subject, known as keywords, are taken mainly from the title and/or sometimes from abstract or full text of the document for the purpose of indexing.

Knowledge Representation : Method of codifying knowledge to enable a computer to store, process and to draw inference from the codified knowledge.

Metadata : In general, it is data about data. Functionally, a metadata is structured data about data, which describes the characteristics of a resource. A metadata record consists of a number of pre-defined elements representing specific attributes of a resource, and each element can have one or more value.

Meta Search Engine : A search engine that simultaneously searches multiple search engines in response to a query.

Meta tag : The HTML element used to demarcate metadata on a Web page.

Natural Language Processing (NLP) : It refers to the area that attempts to make the computer understand natural language.

Ontology : A specification of a representational vocabulary for a shared domain of discourse—definitions of classes, relations, functions, and other objects.

Organising Classification : It refers to the categorization of concepts and their organsation on the basis of genus-species, whole-part, and other inter-facet relationships. It is used to distinguish and rank each subject from all other subjects with reference to its COSSCO relationships.

Parser : A computational process that takes individual sentences or concerned texts and converts them to some representational structure useful for further processing.

Parsing : It refers to the use syntax to determine the functions of words in the input sentences in order to create a data structure that can be used to get at the meaning of the sentence.

Post Coordinate Indexing : Component terms in a compound subject are kept separately uncoordinated by the indexer and the searcher does the coordination of the component terms at the time of searching. Also called *Coordinate indexing.*

Pre-coordinate Indexing : the indexer carries out Coordination of component terms in a compound subject at the time of indexing by following the syntactical rules of a given indexing language.

Process : One of the fundamental categories propounded by Kaiser, which includes mode of treatment of the subject by the author, an action or process described in a document, and an adjective related to the *Concrete* as component of the subject.

Quality of Indexing : The ability to retrieve what is wanted and to avoid what is not wanted.

Relational Indexing : An indexing system developed by J E L Farradane, which involves the identification of the relationship between each pair of terms of a given subject

statement and representation of those relations by relational operators.

Relational Operators : Special symbols used to link the isolates in *Relational Indexing* to create *analets.*

Relative Index : An index showing various aspects of an idea and its relationship with other ideas.

Resource Description Framework (RDF) : A generic framework for describing and interchanging metadata on the Internet. RDF metadata expresses the meaning of terms and concepts in the XML that is understandable to computers.

Role Operators : These refer to a set of notations, which specifies the grammatical role or function of the term which follows the operators and which regulates the order of the terms in an input string in PRECIS. The rules associated with role operators serve as computer instructions for generating index entries, determines the format, typography and punctuation associated with each index entry.

Role : A symbol attached to the index term to indicate the context in which the term has been used.

Search Engine : A searchable database of Internet files collected by a computer program (called a *wanderer, crawler, robot, worm,* and *spider.*

Semantic Net : A directed graph in which nodes represent entities and arcs entities. Arcs are labelled with the names of the relation types—that is, the binary relationship to which the relationship belongs. A single node represents a single entity. Semantic nets can be used to represent various types of knowledge.

Semantics : Semantics is a study of meaning. In an indexing language, semantic relationship refers to the hierarchical and non-hierarchical relationships between the subjects and is governed by *see* and *sees also* references in an index file. Controlled vocabulary serves as the source for *see* and *see also* references.

Semantic Web : The term *Semantic Web,* introduced by Tim Berners-Lee, is the extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The infrastructure of the Semantic Web would allow machines as well as humans to make deductions and organize information. The architectural components include semantics (meaning of the elements), structure (organization of the elements), and syntax (communication).

Software Agent : A computer program that carries out tasks on behalf of another entity. Frequently used to reference a program that searches the Internet for information meeting the specified requirements of an individual user.

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

Spider : A computer program used in indexing and retrieving Web resources with reference to their URLs that contain the given keywords or phrases. A *spider* traverses the Web from link to link, identifying and reading pages. Also called *crawler, robot,* etc.

Standard Generalized Markup Language (SGML) : A non-proprietary language/enabling technology for describing information. Information in SGML is structured like a database, supporting rendering in and conversion between different formats. Both XML and later versions of HTML are instances of SGML.

Subject Directory : A search engine service that offers a collection of links to Internet resources submitted by the site creators or evaluators and organised into subject categories.

Subject Gateway : A subject gateway is recognised to be specialised resources on a particular field and is compiled by the people, not robots. The resources in the subject gateway include Internet catalogues, subject directories, virtual libraries and gateways and these resources are organised into hierarchical subject categories. Also called *Information Gateway.*

Syntax : The grammatical structure consisting of a set of rules that govern the sequence of occurrence the terms in a subject heading.

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

Term Relationship : It refers to the relationship between two or more equally concrete things or phrases in a subject, which may lead to the absence of order of significance and modification of the amplified phrase order as propounded by E J Coates in his subject indexing system. Coates has identified 20 different kinds of relationships by means of prepositions: *of, for, against, with,* and *by.*

Term Significance : It refers to the order of significance developed by E J Coates in his subject indexing system. It states that the most significant term in a compound subject heading is the one that is most readily available in the memory of the enquirer and this leads to the order of significance as *Thing-Part-Material-Action.*

Uniform Resource Identifier (URI) : The syntax for all names/addresses that refer to resources on the Web.

Uniform Resource Locator (URL) : A technique for indicating the name and location of Internet resources. The URL specifies the name and type of the resource, as well as the computer, device and directory where the resource may be found. The URL is a subset of URI. For example, the URL for Dublin Core Metadata Initiative is http://dublincore.org/.

Uniform Resource Name (URN) : A URI (name and address of an object on the Internet) that has some assurance of persistence beyond that normally associated with an Internet domain or host name.

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

Weighting : A method of the allocation of values to indexing terms by using quantitative figures according to their importance in the document.

World Wide Web (WWW) : The panoply of Internet resources (text, graphics, audio, video, etc.) that is accessible via a Web browser. Also called Web or W3.

4.11 REFERENCES AND FURTHER READING

Borko, Harold and Bernier, Charles L. (1978). *Indexing concepts and methods*. New York: Academic Press, 1978.

Chakraborty, A.R. and Chakraborty, Bhubaneswar. (1984). *Indexing: principles, processes and products*. Calcutta: World Press.

Chowdhury, G.G. (2004). *Introduction to modern information retrieval*. 2nd ed. London: Facet Publishing.

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

Gopinath, M.A. (1999). *Indexing process and models*. In: MLIS-03, Block 2 course materials. New Delhi: Indira Gandhi National Open University.

Guha, B. (1983). *Documentation and information: services, techniques and systems*. 2nd rev ed. Calcutta: World Press.

Lancaster, F.W. (1979). *Information retrieval systems: characteristics, testing, and evaluation*. 2nd Ed. New York: John Wiley.

Lancaster, F.W. (1998). *Indexing and abstracting in theory and practice*. 2nd ed. Champaign Illinois: University of Illinois.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4), 309-317.

Prasad, A.R.D. (ed.) (1993). *Artificial Intelligence applications to library and information work*. DRTC Refresher Workshop (May 26-28, 1993). Bangalore: Documentation Research and Training Centre, Indian Statistical Institute.

Prasad, A.R.D. (ed.) (1993). *Semantic web*. DRTC Workshop (December 8-10,

2003). Bangalore: Documentation Research and Training Centre, Indian Statistical Institute.

Quinn, B. (1994). Recent theoretical approaches in classification and indexing. *Knowledge Organization*, 21(3), 140-147.

Ranganathan, S.R. (1964). Subject heading and facet analysis. *Journal of Documentation*, 20, 101-119.

Salton, G. and McGill, Michael I. (1983). *Introduction to information retrieval*. New York: McGraw-Hill.

Sarkhel, J.K. and Mukhopadhyay, P.S. (2002). *Towards a semantic web: a metadata approach*. In: *National Seminar on Information Management in Electronic Libraries* (February 26-27, 2002). Kharagpur: Indian Institute of Technology Kharagpur.

Sarkhel, J.K. (2001). *Information analysis in theory and practice*. Kolkata, Classique Books.

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