CHAPTER C

SUBJECT INDEXING LANGUAGE

- Subject
- Indexing
- Subject Indexing
- Subject Index Entry
- Organisation in Subject Index
- Axioms of Subject Indexing
- Subject Indexing Language and Natural Language
Subject Indexing Language is a language for formulating names of subjects. The term 'Subject Indexing Language' (=SIL) is sufficiently well established as a technical term. The denotation of this term can be well understood by constituent analysis of the compound expression. For that purpose, each of the substantives namely, 'Subject', 'Indexing' and 'Language' would each call for precise definition. The denotation of the term 'Language' has been elaborately dealt with in the previous Chapter B. Hence the terms 'Subject' and 'Indexing' are discussed to bring out the exact denotation of the term 'Subject Indexing Language'.

Cl SUBJECT

To arrive at a precise definition of the term 'subject', it is helpful to get the notions of 'Idea' and 'Information' clarified at the very beginning.

Cl.1 An Idea in one sense, is a unit mental existent, and the term is used to refer to any one of the following:

1 a percept — that is, a meaningful impression of an existent obtained through primary senses;
2 a mental image formed out of association of several percepts — usually called a percept association; and

3 a generic mental image abstracted from several associations of percepts — usually called a concept. Concepts are normally the products of reasoning, generalizing mental operation, mental reflection, imagination and intuition.

Cl.2 For the purpose of communication with others or for self-communication, ideas need to be expressed in terms of symbols of different levels of a language. Usually the symbols are signs, signals, gestures, sounds, pictures, marks, words or their combinations. The language usually includes also the rules for the formation of admissible expressions and their transformations. They are usually used and understood in common by a community. In the case of the English language, the words, phrases, clauses and sentences constitute its symbol system at different levels, as mentioned earlier in section B122. The ideas of different varieties, that is, simple ideas, compound ideas,
complex ideas etc. are expressed in terms of these symbols. Any substantive symbol of a language — natural or artificial — is intended to denote and/or connote an idea; and an idea must have a correlate — perceptual or conceptual, concrete or abstract.

C1.3 The intention of successful communication calls for constructs arrived at by mentally systematizing symbols of different levels of language — natural or artificial.

In communication such constructs are, fully or partially, intended to inspire, urge, advise, warn, appeal, describe, explain, clarify, enlighten, educate, inform, provide recreation or emotional satisfaction, discourage, create emotional disturbance, and so on. A construct mentally systematized with such an intention may be regarded as one having the potentiality of conveying a "message". A message, therefore, is basically the element of intention, as interpreted by the communicatee, underlying one or more such constructs. "Information" is the message conveyed or intended to be conveyed by a systematized body of ideas, or its accepted
or acceptable substitutes (59). Similar notion of "information" may be found in the definition of Brookes: "Information is that which modifies a knowledge structure in any way" (60).

1.4 Regarding any linguistic unit deemed to be a systematized body of ideas conveying a message, or any combination of several such units, two distinct questions can be raised:

1. What is the information content of the linguistic unit(s)? and
2. What the linguistic unit(s) is/are about?

The answer to the first question has to be in terms of sentence or sentences. Generally, non-sentence constructs are not adequate, and therefore are not readily acceptable as answers. On the other hand, the answer to the second question can be in terms of a sentence, or simply in terms of a phrase deemed to be a substitute for a sentence (59) -- "title or title-like construct ... and even un-grammatical forms may not be inappropriate" (61). In other words, non-sentence constructs are adequate, and therefore readily acceptable.
as answers to the second question. Such a construct, used to answer the second question is an "indicative formulation that summarises in its message 'what a particular body of information is about' " (59).

Cl.5  "To tell what is the topic of a play, a picture, a story, a lecture, a book etc., forms part of the 'individuals' mastery of a natural language (and his ability to comprehend). In fact, such activities are often used in language instruction as tests of some aspects of mastery of language... They are the 'starting point' of most requesters when approaching the (bibliographic information) system, or in a dialogue with a librarian or documentalist. They also act as 'personal standards' by which many users judge the relevance of documents produced by the formalized search: reading a document the user explicitly or implicitly applies his above-mentioned ability of "telling what it is about" and compares his result with his starting point" (62). In other words, "what a document is about" is a piece of information, fundamental to the search for documents."
Cl.6 Viewing from this angle, information can be taken to be of two types:

1. Discursive information; and
2. Non-discursive information.

Discursive information is in the form of an answer to questions of the type — "How do you generate a thesaurus using computer?". It would be in the form of a lecture, a discourse, a dialogue, an essay etc.

Cl.7 Non-discursive information is of two types:

1. Quantitative information; and
2. Qualitative information.

Quantitative information is in the form of an answer to questions of the type — "How many wives do you have?". It would be in the form of a numerical fact.

Cl.8 Qualitative information is in the form of an answer to questions of the type — "what are you going to speak about today?"; "what is the 'topic' of your lecture today?"; "what is the 'subject of your lecture today'?". It would be in the form of an indicative formulation that summarises in its message 'what a particular body of information is about'. Exhibit-2
EXHIBIT - 2: TYPES OF INFORMATION
Exhibit 2: Types of Information

- **DISCURSIVE**
  - Answer to questions of the type:
    - How do you generate an information retrieval thesaurus using computer?
      - in the form of a lecture, an essay etc.

- **NON-DISCURSIVE**
  - **QUANTITATIVE**
    - Answer to questions of the type:
      - How many wives do you have?
      - in the form of a numerical fact
  - **QUALITATIVE**
    - Answer to questions of the type:
      - What is the subject of your lecture today?
      - in the form of an indicative summarizing statement.
gives these types of information in the form of a diagram.

C1.9 On the basis of the analysis and explanation given above, it has been stated (59) that

"A subject is essentially a piece of non-discursive information and it is conveyed by an indicative formulation that summarises in its message what a particular body of information is about".

C2 INDEXING

'Index' has been defined in the British Standard (63) as "A systematic guide to the location of words, concepts or other items in books, periodicals or other publications. An index consists of a series of entries appearing, not in the order in which they appear in the publication, but in some other order (ex: alphabetical) chosen to enable the user to find them quickly, together with references to show where each item is located". According to Wheatley, "An index is an indicator or pointer out of required information" (64). Welisch (65) has investigated the term 'index', its etymology, semantics, pragmatics, derivations and its relation to other concepts, based on listings of usage.
C2.1 The unit of an index is the entry, which is a record specifying an existent along with its address. An entry consists of index terms. An index term is a word or term such that

1. it has a definite meaning;
2. it cannot be decomposed into two or more meaningful units; and
3. it is capable of being manipulated independently of other index terms (66).

NOTES The index terms may be names of persons, names of corporate bodies, project codes, terms denoting subject etc.

Any set of methodically arranged entries is an index.

The process of preparing an index is indexing.

C3 SUBJECT INDEXING

In the context of subject indexing, an entry always pertains to a source of information or document. In such a case, the role of indexing is to completely specify the subject of the document or source of information being indexed, using index terms and their
relations by means of order-preserving transformations (66). In other words, accurate indexing (subject indexing) preserves the order of the subject in the order of the index terms reflecting their interrelations, to specify completely and exactly the subject of the document being indexed. The order of the subject is the same as "the measure intension" of Ranganathan. "Intension of a subject has for its measure the number of characteristics used in deriving it from the original universe" (67). Subject Indexing: Language is the language used for the complete and exact specification of a subject, using as its symbols both the index terms denoting subjects and their relation indicators, according to a grammar prescribing an order of the symbols, reflecting the order of the subject.

C3.1 In order to identify the index terms and their interrelations to specify the subject of the document being indexed, a technique called subject analysis or document analysis is made use of in SIL. As mentioned earlier in section B7, a SIL may use a vocabulary control tool indicating hierarchy, such as a thesaurus, for selection of appropriate index terms.
(including broader terms, and narrower terms) to preserve the order of the subject being indexed. A SII may also propose a set of semantic categories for its index terms which would be referred to in its rules of syntax for formulating statements specifying subjects of documents. In order to indicate the relation among the index terms, a SII may also have a set of relation and/or role indicators.

C4 SUBJECT INDEX ENTRY

An entry specifying a subject along with its address is a subject index entry. It consists of index term(s) denoting a subject. Index terms denoting a subject are generally constituents of the name of the subject and are the keywords of the subject.

C41 Functions of Subject Index Entries

Subject index entries in general perform three functions. They are as follows:

1 Locating function — to permit the location of index entries for the subject sought. The subject analysis may reveal five constituents in the name of the subject for a document. Subject index entries may be made
Each with all the five concepts, but the number of entries need not necessarily be five. Entries would be made and arranged in the places for only the index terms through which the subject is likely to be sought. It may be just three. This is actually according to the prescription of the Canon of Sought Heading (69).

2 Comprehending function — to give data for the comprehension of the index entries to permit relevance prediction. The content of the index entry providing the context of the index term is given to facilitate searchers to comprehend the full meaning such that a judgement may be made — at least a provisional decision as to whether it is worth seeking fuller information.

3 Organising function — to aid the location of entries for subjects related to the one being sought. Under an index term, the sub-arrangement of entries in a way that brings all related subjects together. This is also done by providing
Cross Reference Entries, directing the searcher to look under other related index terms, using See and See also entries.

C42 Components of Subject Index Entries

A subject index entry comprises a heading at its beginning called Lead Heading. The index term occupying the first position in a Lead Heading is called the Lead Term. The potency to decide the position of an entry among the various entries in a subject index is concentrated on the Lead Heading and mainly on the Lead Term. In other words, the Lead Term mainly decides the location of the entry in the index. The Lead Term is the basic constituent of an entry catering to the locating function mentioned in section C41 above. The Lead Term is also called as an Approach Term or Access Term. The other occupants of a Lead Heading may consist of other index terms (index terms with or without other auxiliary words/function words forming a phrase) ‘qualifying’ the Lead Term making its meaning clear.

C42.1 A subject index entry may also contain what is called a 'Context Heading' as an occupant of
the immediate next line to the line occupied by the Lead Heading. A Context Heading consists of index terms with or without other auxiliary words providing the context to the Lead Heading, making the meaning of the Lead Heading still more clear. Generally the Context Heading aids the searcher to comprehend the meaning of the entry and helps him in predicting the relevance of the entry to the subject sought by him. In other words, Context Headings are provided in a subject index entry to cater to the comprehending and relevance prediction function mentioned in section C41 above. Context Headings to the same Lead Heading form sub-entries and are also used for systematic grouping by bringing together subjects related to the subject denoted by the Lead Heading. In this sense Context Headings are also used to cater to the organizing or relating function referred to in section C41 above. The Context Heading needed to make precise the meaning of a Lead Heading for index language use is less than for ordinary language use. Also it has been observed that ‘many words are capable of independent use, especially in book indexes’ (69).
It has also been observed in Aberystwyth index language test that "the provision of context (Context Heading) for search screening (relevance prediction) in either its normal pre-coordinate form or as a file for use during post-coordinate searching showed a strongly beneficial effect in suppressing non-relevant (entries) for a small recall loss" (70). Another feature of the Context Heading is concerned with how much of context is desirable in an index entry. "The rule of specificity is sometimes interpreted to mean that an index term (entry) should be coextensive with the theme (subject) of the item (document) being indexed; that is, the complete subject analysis of the item (document) should be expressed in a single entry" (71).

C42.2 Apart from using the Context Headings to a Lead Heading to bring together subjects related to the Lead Heading, other entries called Cross Reference Entries or Directing Entries are used in subject indexing to cater to the relating (organizing) function referred to in section C41 above. These entries direct the user from one Lead Heading (referred from heading) to another Lead Heading (referred to heading).
In between the 'referred from' and 'referred to' headings, directing elements such as 'See' or 'See also' are used. If the searcher will not find any entry under the 'referred from' heading, then the directing element 'See' is used to direct him to refer to the 'referred to' heading under which entries may be found. These 'See entries' refer the searcher from a non-standard (synonymous or variant-form of the) heading to the standard heading. If the searcher is likely to find entries under another heading which is related to the heading being referred to by him, the directing element 'See also' is used. Generally, the 'referred to' heading denotes subjects related to the one denoted by the 'referred from' heading hierarchically (broader or narrower) and/or associatively.

C42.3 Subject index entries other than cross reference or directing entries contain the address or location of the source of information or document. The address or location may be an index number or reference number or location number or accession number or a record number of the record containing
full bibliographical details of the document, sufficient for identifying uniquely the document being indexed. Instead of the index number or record number the full bibliographical detail may be given in the subject index entry itself in place of the address or location.

C43 Subject Index Heading

A heading in a subject index entry denoting the name of the subject is a subject heading. It consists of index terms (generally in noun form) which are the keywords of the subject. It may also consist of other auxiliary words or function words such as, prepositions, conjunctions etc. These auxiliary words are used in the heading to make the heading read like a grammatical phrase which may permit quicker and more accurate comprehension for relevance prediction purposes. But these would affect the sub-arrangement of index entries. These auxiliary words or function words or relational connectors are used to symbolise relations between index terms in order to clarify the meaning of the heading and for unambiguous representation of the subject. These relational connectors are of various kinds. Prepositions
are used as 'between — index term 'relational connectors in Articulated Subject Index (72).
British Technology Index uses a limited set of punctuation marks (73), as an alternative to auxiliary words to which special meanings are assigned for the purpose as relational connectors.
Farradane (74) uses pairs of punctuation and related symbols available in the standard typewriter keyboards.
Mimeur (75) has proposed a fuller range of symbols, variously logical or semi-pictorial in character.
In all these cases users who want a clear understanding (which may not always be necessary) have to learn the convention.

Order of Index Terms

In a subject heading combinations of index terms without relation indication, and even combinations of terms joined by relation indicators which do not give the direction of relation, can yield multiple meanings. The subject heading 'Towers Erection Cranes' could mean "the erection of towers using cranes" or the "erection of cranes using towers" (73).
Similarly, the subject heading 'Bacteria Destruction Paint' could mean "destruction of paint by bacteria"
or "destruction of bacteria by paint" (76). The subject headings could be made unambiguous using appropriate relational connectors or role indicators in the form of auxiliary words as shown below:

- Towers Erection (by)/(using) Cranes
- Towers Erection (of) Cranes
- Bacteria Destruction (by)/(using) Paint
- Bacteria Destruction (of) Paint

The alternative to this is to dispense with relational connectors altogether and rely on index term order (sequence) to indicate relations. This again requires the user to understand a convention, but the convention is one closely corresponding to natural language, being either the order in which concepts are given in a simple sentence (with active direction relations) or the reverse of simple sentence order (with passive direction relations). Reverse sentence order i.e., passive order, is rather commoner than direct sentence order. One underlying principle of index term order is that of contextual dependency.
which is the same order as that prescribed by Ranganathan's Wall-picture principle and other principles for facet sequence derived from it (45, 46, 57). Increasing or narrowing contextual order is passive order, increasing or widening contextuality is active order. The index term order is also known as citation order. If the index terms given in the subject headings above are arranged in the order 'Actant-Action-Actor' then each of their denotation becomes unique. The order 'Towers, Erection, Cranes' would represent "erection of towers using cranes" and the order 'Cranes, Erection, Towers' would represent "erection of cranes using towers". Similarly the index term order, 'Bacteria, Destruction, Paint' would represent "destruction of bacteria by paint" and the order 'Paint, Destruction, Bacteria' would represent "destruction of paint by bacteria". Costes (73) has observed that "A mixed system using both connectors and position indication of relation, enables one to economise on the number of relators needed".
C5 ORGANIZATION IN SUBJECT INDEX

Structure in an index provides the capacity for organization. If it is to serve its function efficiently — to organize and provide access to a collection — it must reveal its logical organization as quickly as possible to users (78). In the arrangement of subject index entries the main principle followed is the alphabetical sequence. Hence providing an organizing sequence in alphabetical arrangement is almost impossible.

C5.1 On the other hand, a subject indexing language is used to summarise in indicative formulations what the contents of a source of information are about. The purpose of these summarizing indicative formulations is to create groups of sources of information to facilitate expeditious retrieval of information about them by providing necessary and sufficient access points. By implication, therefore, a SIL is a classificatory device; and in that sense, a classificatory language. The classification forming the basis
may be either organizing (systematic) or associative or a combination of both. To serve the whole purpose of exhaustive retrieval, an organizing classification must be complemented by an associative classification or vice versa. An organizing classification can serve as the source for deriving associative classification (79), as in the case of Colon Classification and the Chain Procedure. Provision of organizing and associative classifications is characteristic feature of SIL. Provision of organizing classification effect in alphabetical arrangement is a striking feature of a particular SIL based on 'Deep Structure of SIL', which is discussed in Chapter C6.

C6 AXIOMS OF SUBJECT INDEXING

Ranganathan's contributions to the field of universal classification, Calvin Mooers’ contributions to the topological representation of molecular structured and the techniques of representing data structures for computer manipulation, have all contributed to the development of a system of representation of concept relations (80), and to the formulation of axioms of subject indexing (81-83).
These axioms viewed from the angle of an enquiry or search of a subject index are given below. These axioms have been applied in comparing recent computer-based subject indexing systems discussed later in Chapter I.

C61 Axiom of Definability

The axiom of definability has been stated as follows: "The compilation of information relevant to a topic can be delegated only to the extent to which an enquirer can define the topic (subject) in terms of concepts and concept relations".

C61.1 A large part of the information required by an user consists of a definable part and an undefinable part. Consider for example the following query: "What is the best synthetic route to substance X?". What an enquirer, who wants an answer to the above query, can demand of a bibliographic information system is merely the documents on "comparison of techniques for synthesis of substance X". The choice of comparison criteria and the judgement of their weights relative to one another are too subjective, and hence the selection is left to the enquirer.
In order to define the topic in terms of concepts and concept relations, the question that has to be asked is "what the particular enquiry is about?". The answer to such a question (the enquiry is about the "comparison of techniques for synthesis of substance X") yields the concepts and concept relations that are to be searched for. In other words, the concept of "aboutness" (84, 85) is fundamental to subject indexing in particular, and to bibliographical information systems in general.

C62 Axiom of Order

The axiom of order has been stated as follows: "Any compilation of information relevant to the topic of an inquiry is an order-creating process".

C62.1 Subject indexing is a process of transformation or mapping of the name of the subject in an indexing language, as mentioned in section C3. The purpose of subject indexing is to create groups (organization) of sources of information to facilitate expeditious retrieval of information. If the entries to an inquiry are isolated or at least localised at a
predeterminate place the enquirer is relieved of scanning the entire set of entries. This is facilitated if the entries are arranged in an organized way. Ranganathan has named this arrangement APUPA arrangement (86). If the focal point of interest of the enquirer is taken as the Umbral Region, he would like to have fanned out on either side of the Umbral Region, subjects having successively a decreasing bearing on the Umbral subject. The two regions on either side of the Umbral Region are called his Penumbral Regions. The Penumbral Regions will ultimately thin out into the Alien Regions on either side. Such an Alien - Penumbral - Umbral - Penumbral - Alien (APUPA) arrangement of subject index entries will give the enquirer greatest satisfaction as he can concentrate his attention on a limited, isolated section of the subject index. Such an arrangement in the alphabetical sequence of subject headings is not possible to achieve. Approximation to such an organizing sequence is necessary in subject indexes.

C63 Axiom of Sufficient Degree of Order

The axiom of sufficient degree of order has been stated as follows: "The demands made on the degree of order increase as the size of the collection and/or the frequency of the searches increase".
Subject indexing should be so precise and deep enough to distinguish clearly, even minute differences in the subjects of documents indexed. This is similar to the concept of "resolving power" mentioned by Ranganathan (87). The demands made on the degree of order increase as the size of the sources of information indexed increases. In the early stages, any bibliographic information system will perform well because of the small size of the collection. If the indexing system does not possess the necessary resolving power, then as the size increases, a larger number of documents would be indexed as belonging to the minutest subject specification attainable by it. The size of each individual group would increase enormously if the resolving power of the indexing system is small. As a consequence, the subject indexing system should have the capacity to represent coextensively the subject of the document being indexed. This aspect of coextensive or complete and exact representation of the subject of a document is the concern of the axiom of fidelity discussed below.
The axiom of fidelity has been stated as follows: "The accuracy of any directed search depends on the fidelity with which concepts and statements are expressed in the search file."

The accuracy of a search is strongly dependent on how precisely (completely and exactly) the subjects of the documents are represented in the subject index. It is also dependent on how precisely the enquirers' topic can be expressed in the indexing language. Fidelity implies coextensiveness — "the representation in the subject heading of the measure of incidence of each of the relevant characteristics of the subject embodied in the document indexed" (88). Fidelity or coextensiveness depends not only on the vocabulary — the accuracy of the index terms, the specificity of the index terms and the resilience to accommodate new terms, but also on the expressiveness of the grammar, particularly on the syntax of the grammar of the indexing language.
The axiom of predictability has been stated as follows: "The accuracy of any directed search for relevant information depends on the predictability of the modes of expression for concepts and statements in the search file."

Natural language provides a large variety of modes of expression for a concept or statement. In order to search the subject index file and to avoid loss, the searcher should know in advance the different modes of expression in which the topic of enquiry is represented in the index. In other words, one must be able reliably to construct or predict by which modes of expression a topic in question is represented in the subject index file.

If accurate searches for general concepts and statements are demanded then author lingual expressions are inadequate because of the unpredictability of their modes of expression. If new and different modes of expression for one and the same subject is avoided in the subject index from the very beginning, a higher degree of predictability is attained. Predictability depends on consistent representation of concepts using index terms and their interrelations using the grammar of the subject indexing language.
C7 SUBJECT INDEXING LANGUAGE AND NATURAL LANGUAGE

There is a distinct difference between a SIL and a natural language, especially from the point of view of users of a SIL. The specific purpose of grouping by a SIL calls for a grammar of its own quite distinct from that of any natural language in general. The communicator in a natural language as well as his communicatee have to be well acquainted with the grammar of the natural language concerned, for the sake of successful communication. On the other hand, the formalator of subject headings alone has to be thorough with the grammar of the SIL concerned. The user (searcher) of subject headings (communicatee) is expected to be least concerned about the grammar of the SIL. The subject headings are expected to be formulated in such a way that

1 on the one hand, they should serve all or a selection of their respective functions; and 
2 on the other, the users with their knowledge of natural language and common sense should be able to understand each of them.

From the point of view of the user, a formally structured subject heading is a deviation from the regular structure of his natural language; but it is comprehensible to him.
provided it admits of the right interpretation by analogy with some logical structure known to him. Some kind of knowledge of logical forms, though with most users is not explicable, is involved in all understanding of subject headings (89).

C7.1 Obviously, to satisfy these conditions, and to be efficient and effective, a SIL has to take extraordinary burden which gets reflected in some extraordinary features. One such feature is that for every query in an user's mind, a SIL is expected to lead him to a point where it displays a pattern which, among various purposes, serves as a mechanism to "enlighten" him instantly about the logical form he needs to know for the understanding of subject headings. The 'terms and their relations' of a query may deviate from those of the subject headings of a SIL, but a SIL must entertain the query and provide aids to reformulate it according to its formal structure.

C7.2 The user knows a set of terms. With a particular term in mind he approaches the subject index. The features of the index have led him to the match. At the match-point, the user sees the subject headings, many of which are combinations of terms. The term combinations display a regularity which the user deduces instantly; and thus he learns the grammar of the SIL.
concerned, in his own way. With this knowledge he explores the set of subject headings, having his matched term in common; he selects those which he feels to be relevant to his information need. The subject headings also provide him with other Leaf Headings through cross references. These referred to headings may be relevant to his need though they do not contain his matched term. After the purpose is served, the user can easily afford to forget the "impression" of the grammar of the SIL that he gained, for, the SIL takes the responsibility of offering him the necessary "impression" instantly as and when he consults the subject index. In this sense, a SIL differs from a natural language considerably. This grammar, based on the "Deep Structure" of Subject Indexing Languages is discussed in the next Chapter.