Annals of Library and Information Studies 52, 1; 2005; 15-24

SPECIATOR BASED FACETED DEPTH CLASSIFICATION'S APPLICATION IN THESAURUS CONSTRUCTION

Rajendra Kumbhar
Reader
Department of Library and Information Science
H. P. T. College
Nasik, Maharashtra.
E-mail: kumbharrajendra@yahoo.co.in

The article narrates the experience of constructing a thesaurus of LIS terms by using speciator based faceted depth classification schedule. It also explains the advantages of this method in establishing various thesaural relationships.

INTRODUCTION

Communication of information is one of the basic necessities of human societies of all times. Information in any format can be communicated through various means across space and time. Effective organization of information sources is as important as its communication. This task of effective organization of information sources is entrusted to libraries and information centres.

While organizing information sources in systematic order, libraries and information centres create two sets of databases. The first set consists of actual information sources, while the other consists of entries describing information sources contained in the first set. The first set of databases (i.e. the documents) is usually arranged by using the call numbers, prepared according to a classification scheme. While describing the information sources in the second set, usually, three types of entries are prepared. These are author, title and subject entries. The subject entries in the second set are arranged under the terms representing the subject contents of the information sources. This set of subject entries serves as subject index to the first set. The set of entries representing the subject of whole document is usually referred to as subject catalogue. Whereas, the set of subject entries referring to parts / contents of documents, such as articles in periodicals, is called subject index. In the field of library and information science (LIS), the word 'index', without any epithet mostly stands for subject index.

The steps involved in preparing both, the subject catalogue and index are same and are as follows:

a) Analyzing contents of document and to identify its subject,

b) Translating the subject name into a standard term/s or phrase/s and preparing entries under the standard terms/s, and

c) Arranging the subject entries in a systematic order.

The standard terms used in the index to represent the subject of the document are called index terms. They are also called as descriptors because they describe the subject of the document.

The set of index terms along with rules for their selection and coordination is referred to as, 'indexing languages'. The appropriateness of index terms is vital because they represent subject of both, the document and the reader's request. Thus, using the indexing language, information is stored and retrieved.

Indexing languages are categorized into natural language (also called uncontrolled language) and controlled language [1]. Natural language consists of words taken from the title, abstract or text of the document. Natural language offers a great variety of terms in the form of synonyms, antonyms, homonyms, adjectives, etc. Thus, natural language is complex and rich in variety. These features of natural language create number of problems in information storage and retrieval (ISR). Controlled languages are those in which both the terms that are used to represent subjects and the process by which the terms are assigned to particular document's description are controlled. Such a control of terms is known as vocabulary control. For controlling vocabulary in ISR various tools are used.

Thesaurus is one of the prominent tools used in vocabulary control. This tool has to be deliberately constructed for a specific subject or group of subjects. In constructing a thesaurus, faceted depth classification schedules can be used. This article narrates the experience of constructing a thesaurus of Library and Information Science (LIS) terms. The thesaurus was constructed by using a specially constructed speciator based faceted depth classification schedule of LIS.

DEPTH CLASSIFICATION SCHEME

Depth classification is 'a scheme of classification fitted to reach co-extensiveness and expressiveness in the classification of micro thoughts having many rounds and levels of facets and isolates of high orders in any or all of them [2].

SPECIATOR

Speciator is an isolate idea used as a qualifier for an isolate or another speciator. When a sub-isolate i.e. a qualifier to an isolate, has a possibility of going with more than one isolates, it is listed as speciator, instead of array division, e. g. the sub-isolate 'computerized' can go with the isolates such as cataloguing, classification, circulation, etc. As such the sub-isolate 'computerized' is listed as a speciator, only once, and can be combined with any appropriate isolate as and when needed to form class number for a compound subject.

There are two kinds of speciators. Speciator kind 1 (Sp1) qualifies an isolate idea. The connecting symbol 'hyphen' (-) is used to connect the speciator kind 1 to an isolate. Speciator kind 2 (Sp2) qualifies the speciator kind 1. Speciator kind 2 is attached to speciator kind 1 by using the connecting symbol 'equal to' (=).

For constructing the present LIS thesaurus, first a depth classification schedule of LIS terms was constructed. The methodology adopted in constructing both the depth classification schedule and the thesaurus is described herewith.

METHODOLOGY ADOPTED IN CONSTRUCTING THE DEPTH CLASSIFICATION SCHEDULE OF LIS

The depth classification schedule was constructed by adopting blending method suggested by Ranganathan [3]. The steps adopted in constructing the depth classification are:

Understanding the subject

For knowing the highways and byways of LIS a large number of Indian and foreign documents were browsed and read.

Identification of the sample population

For constructing the depth classification schedule, 500 assorted micro documents were analyzed. This was done to have representative sample and to avoid duplication. Only the micro documents were analyzed, as they deal with the subject in depth. A schedule based on micro documents ultimately proves helpful in classifying macro documents also. The micro documents analyzed were articles published in various documents. An illustrative list of the documents from which the micro documents were analyzed are:

- Full text periodicals available in the hard copy format: *Journal of Documentation*, *International Cataloguing and Bibliographic Control*, *Knowledge Organization*, *Library Quarterly*, *Library Trends*, *Program*, *IASLIC Bulletin*, *ILA Bulletin*.
- LIS periodicals available on the Internet
- Seminar, conference, festschrift and workshop volumes: ILA, IASLIC, DRTC, and CALIBER.
- Abstracting periodical: LISA.

The sample selection was such, so as to have adequate representation to both Indian and foreign literature. In order to satisfy the principle of literary warrant [4] the actual micro documents were analyzed. To satisfy the Canon of Currency [5] the literature published during the year 1994 to 1999 was analyzed.

Analysis of data

As far as possible the full text of the micro document was read, so as to have sufficient understanding of the contents. However, to have wider coverage, some data was collected from the abstracting sources also. The micro documents were analyzed by applying Ranganathan's fundamental categories [6]. During the analysis, the terms were identified as isolates belonging to various facets and their levels, i.e.,

- First round first level personality [1P1]. First round second level personality [1P2]
- First round first level matter property [1MP1] and first round second level of matter property [1MP2]
- First round of energy [1E] and second round of energy [2E]

While analyzing the micro documents the terms representing speciator of kind 1 (Sp1) and speciator kind 2 (Sp2) were also identified.

Micro documents were analyzed to get isolates going with the first three facets of the PMEST, i.e. Personality, Matter property and Energy. These three facets always represent special isolates going with different main classes. The other two facets i.e., space and time were not considered while analyzing the micro documents because their isolates lead to common isolates going with more than one main classes.

Recording of terms

Four by six-inch cards were used to record the results of the analysis. Each card contained following information:

i) Serial number in the right hand top corner.

ii) The subject string or the feature heading: This included the name of main class, i.e. library science, isolate and speciator terms. The connecting symbols of the respective facet was used to separate isolates of each facet, whereas the hyphen (-) and the equal to (=) punctuations separated the (Sp1) and (Sp2) respectively.

iii) The focus: In this part the bibliographical details of the micro documents were recorded. This included – name of the author (s), title of the micro document, host document along with its volume number, issue number, date and page numbers.

iv) Annotations: Whenever the title failed to express the subject contents of the micro document an annotation was added to decipher the subject of the micro document.

There were some micro documents, which contained information on more than one subject. For such micro documents more than one entries were prepared that is how for 500 micro documents 590 entries were prepared.

Grouping of isolate terms

The isolates belonging to different facets were scattered throughout the 590 entries. These isolates were noted down facet-wise.

Grouping of speciators

The speciators of each facet were first grouped under the Quasi Isolate i.e. an isolates which is not true isolates but a name given to group of speciators.

Arrangement of isolates and their speciators

The isolates of various facets and their speciators were arranged by applying Principles of Helpful Sequence [7]. Help of various Devices and Canons [8] was also taken to arrange the isolates and the speciators.

Assigning notations

The present depth classification schedule was constructed as an extension of the Library Science main class '2' in the seventh edition of Colon Classification (CC). As such the notational system of CC was adopted in the present depth classification scheme. Notations to the ranked isolates and speciators were assigned by considering the sectors available and the frequency of occurrence of isolates and speciators.

THESAURUS

Derivationally the word thesaurus originated from the Greek language and symbolises the concept of treasury or store- house of knowledge. The Webster's dictionary has defined thesaurus as, 'a useful literary collection or selection especially a book of synonyms and antonyms'. The UNESCO has defined the thesaurus as, a 'vocabulary of controlled indexing language formally organised so that the *a priori* relationships between concepts (e.g., 'broader' and 'narrower') are made explicit' [9]. This definition lays emphasis on *a priori* relationships, which are document independent. The British Standards 5723 defines a thesaurus as, 'a means of displaying the terms in a controlled indexing language, together with indications of their apparent relationships' [10]. Ajoy Kumar Roy provides definition, giving list of those whose natural language expressions are controlled. He defines thesaurus as, 'an IR thesaurus is chiefly a terminological control device for transformation of natural language (NL) expressions, used by authors, referees, publishers, indexers and users who form the various links in the information transfer chain into a more constrained system of vocabulary' [11]. Thesaurus can also be defined functionally and structurally. Functionally a thesaurus is a terminological control device used in translating the natural language into the systems language. Structurally it is a tool consisting of hierarchical controlled set of terms linked by hierarchical or associative relations, which mark any needed equivalence relations (synonyms) with terms from natural language and concentrate on a particular area of knowledge [12].

While emphasizing the standardization function Rowley [13] defines thesaurus as, 'a compilation of words and phrases showing synonyms and hierarchical and other relationships and dependencies, the function of which is to provide a standardized vocabulary for information storage and retrieval systems'.

Identifying the systems which make use of thesaurus, Aitchison, Gilchrist and Bawden [14] defined thesaurus as, 'a vocabulary of controlled indexing language, formally organized so that *a priori* relationships between concepts are made explicit, to be used in information retrieval systems, ranging from the card catalogue to the Internet'. The last word in this definition indicates the potential usefulness of thesaurus in Internet information retrieval.

Language thesaurus and information retrieval (IR) thesaurus

The various thesauri available can be categorized in two broad categories, i.e. the language thesaurus and IR thesaurus. The language thesaurus is primarily a dictionary of synonyms, though structurally it may differ, to some extent, from ordinary word dictionary. Roget's thesaurus [15] is a classic example of language thesaurus, which, according to Vickery [16] is a tool for writing in English. The IR thesaurus collects terms of a subject/s and organises in such a manner so as to bring out the relations between concepts. Thinking on the same line Schultz, quoted by Lancaster, [17] expresses that the purpose of Roget's thesaurus is to give an author a choice of alternative words to express one and the same concept whereas the IR thesaurus tends to be prescriptive.

In any information storage and retrieval system (ISRS) the indexing plays an important role. An index describes subject of document with the help of descriptors arranged in systematic order. The indexer and the user are the two main parties who have to use precise terms in indexing and searching respectively. However, both of them in the process of information storage and retrieval (ISR) always comes across variety of terms (vocabulary) for representing one and the same concept. Such a variety of terms must be controlled and standardized for having expected recall and precision. Such a controlling of terms in ISR is known as vocabulary control. Vocabulary control is warranted for two reasons -

1. To promote consistent representation of subject matter by indexers and searchers, thereby avoiding the dispersion of related materials, through control (merging) of synonymous and nearly synonymous expressions and by distinguishing the homographs.
2. To facilitate the conduct of a comprehensive search on some topic by linking together terms

whose meanings are related paradigmatically or syntagmatically [18].

Davis and Rush [19] explain the concept of vocabulary control in the following words.

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

METHODOLOGY ADOPTED IN CONSTRUCTING THE THESAURUS OF LIS TERMS

A thesaurus can be constructed by adopting the *a priori* or the posteriori method [20]. Each of these methods has its own merits and demerits. The present thesaurus of LIS terms was constructed by adopting the blending method; so as to avail the advantages of both the methods.

Sources for thesaural terms

The thesaurus contains descriptor (preferred terms) and non-descriptor (non-preferred terms) terms. The newly constructed depth classification schedule formed the basic source for descriptor terms in the thesaurus. The entries in the thesaurus were directly typed in the MS WORD file. Subject terms going with following three different components in the depth classification schedule of LIS, formed the descriptors.

Isolate terms alone, e.g., ACADEMICLIBRARIES, CIRCULATION,

Speciator terms alone, e.g., the speciator terms from (Sp1) under the (QI) 'by information storage and retrieval properties' of [1MP1] facet independently qualified as descriptors e.g., CALL NUMBERS, NOTATIONS. Similarly, speciator terms from (Sp1) under the (QI) 'by tools', 'by influencing factor' and 'by method' of [1E] and [2E] facet qualified as independent descriptors, e.g., SYSTEM ANALYSIS, DIGITIZATION, etc.

Combination of speciators with isolates, e.g., (Sp1) 'Junior' combined with the (1P2) isolate 'Cataloguer' forms the descriptor 'JUNIOR CATALOGUERS'.

The non-descriptor terms were gathered from the micro documents and other sources like glossaries and encyclopaedias. These were added to the computer file at appropriate places while typing the descriptors.

Compound terms: Ideally, any concept in the thesaurus and thereby in the process of ISR should be represented by using single term. That is why the UNESCO's *Guidelines* as well as almost all the manuals on the thesaurus construction recommend factoring of terms. The *Guidelines*, however, mentions that factoring is not mandatory. In spite of such a suggestion for representing a subject with single term, the nature of subjects discussed in the documents does not always allow do so. Complex subjects need to be expressed by combinations of two or more terms. The terms formed by combining two or more terms are called compound terms. Compound terms increase specificity and ultimately enhance performance of ISRS. Considering this aspect, the UNESCO's *Guidelines* prescribed the use of compound terms in thesauri. So in the present LIS thesaurus also lists terms in compound form.

The classification schedule which formed a basic source for thesaural terms, was developed by applying the concept of speciator. The speciator terms were combined with the isolate terms to form compound terms. Illustrative examples from each facet are:

a) The (Sp1) 'Automated' was combined with the [1P1] isolate 'University library' to form the compound term 'AUTOMATED UNIVERSITY LIBRARIES'.

b) The (Sp1) term 'Rural' was compounded with the [1P2] isolate term 'User' to form the compound term 'RURAL USERS'.

c) The compound term 'COMPUTERIZED SDI' was formed by combining the (Sp1) term 'Computerized' with the [1MP1] isolate term 'SDI'.

d) The (Sp1) term 'ISO' was compounded with the [1MP2] isolate term 'Standard' to form the compound term 'ISO STANDARDS'.

e) The compound term 'QUANTITATIVE EVALUATION' was formed by combining the (Sp1) term 'Quantitative' with the [1E] isolate 'Evaluation'.

f) The (Sp1) term 'Security' was combined with [2E] isolate term 'Problems' to form the compound term 'SECURITY PROBLEMS'.

In addition to these, compound terms were also formed by combining two speciators with each other, for example, RARE BOOK WEBSITES. This was done considering the literary warrant. These terms were combined because factoring them may lead to loss of meaning.

Compound terms were also formed (though very few) by combining the two isolates across facets, e.g., the [1P2] isolate 'Personnel' was combined with the [1E] isolate 'Management' to form the compound term 'PERSONNEL MANAGEMENT'. Such combinations were formed only for those terms warranted by the literature.

Combinations of terms have been formed primarily for those subjects which, occurred in the analyzed micro documents, i.e., based on literary warrant. Some compound terms were also formed on the empirical basis. For forming such compound terms help of Harrod's Glossary was taken.

THESAURAL RELATIONSHIPS

Methods used for establishing the three types of thesaural relationships are:

The equivalence relationship

The data required for establishing the equivalence relationships was collected during the analysis of micro documents. Additional data required for this purpose was gathered from glossaries and encyclopaedias. This relationship has been shown in the thesaurus through the usual convention of USE as a prefix to the preferred term and UF (USE FOR) as a prefix to the non-preferred term. Reciprocal entries have been made from preferred to non-preferred term and vice-versa. The equivalence relationship existed between the terms of various categories such as:

Synonyms

The equivalence relationship was established between the preferred term and its non-preferred equivalent, e.g.,

LIST OF ADDITIONS

UF Accessions lists

A reciprocal entry has been prepared under the non-preferred term, e.g.,

Accessions list

USE LIST OF ADDITIONS

Terms of different linguistic origin

Different linguistics origins may have different names for the same concept, e.g., periodicals are also referred to as serials. Relations between such variations have been established giving reference from preferred to non-preferred term and vice-versa, e.g.,

PERIODICALS

UF Serials

Serials

USE PERIODICALS

Acronyms

As recommended by the UNESCO's *Guidelines*, the well-known acronyms were used as preferred terms and cross-reference was given to its full form, e.g.,

KWIC

UF Key-Words-In-Context

Key-Words-In-Context

USE KWIC

The classification schedule does not help in establishing the equivalence relationship.

'circulation' belongs to [1MP1] facet, whereas the type of libraries i.e. the 'academic libraries' belongs to [1P1] facet. As the two isolates belong to different facets the associative relationship between them was established, i.e.,

ACADEMIC LIBRARIES

RT CIRCULATION

The associative relationship too is reciprocal, so a reciprocal entry was prepared under another related term i.e.,

CIRCULATION

RT ACADEMIC LIBRARIES

Some times a document has isolates coming from the more than two facets also. In such cases the associative relationship was established between all of them, giving reciprocal reference from each other, e.g.,

The subject, 'designing [1E] standards [1MP1] for academic libraries [1P1]', derives associative relationship such as:

ACADEMIC LIBRARIES

RT DESIGNING
STANDARDS

DESIGNING

RT ACADEMIC LIBRARIES
STANDARDS

STANDARDS

RT ACADEMIC LIBRARIES
DESIGNING

Speciators under the (QI) 'by cause' to [1MP1] isolates

The speciator to [1MP1] under the (QI) 'by cause' represented associative relationship. This is because these speciators fulfill the criteria of 'cause – effect', recommended by the UNESCO's *Guidelines* for establishing associative relationship.

Speciators under the (QI) 'by tools' to [1E] and [2E] isolates

It has been noticed that the speciators to [1P1], [1P2], [1MP1] and [1MP2] (except the speciators under the (QI) 'by cause') represent a type, part or a class. However this is not true of all the speciators to [1E] and [2E]. All the speciators to these facets do not represent types, parts or class of isolates. So all of them do not form part of the associative relationship. As such a different treatment is given to them while establishing thesaural relationships for terms prepared by combining these speciators. Illustrations of these treatments are:

Speciators under the (QI) 'by tool', 'by influencing factor' and 'by methods' have been shown as RT to the respective isolates in [1E] and [2E] facet. That is because they fit in the 'operation and instrument' and other criteria suggested by the UNESCO's *Guidelines* for forming associative relationship, e.g., the speciator 'fill rate' forms RT to the descriptor 'evaluation'. This was done for two reasons. (a) the speciator 'fill rate' does not represent type of evaluation, and (b) it fits in the type of associative relation named 'an operation or process and its agent or instrument', given in the UNESCO's *Guidelines*.

Terms within a facet

It is usually noticed that most of the terms indented under an isolate in a faceted systematic display form hierarchical relationship. However as warned by Aitchison, Gilchrist and Bawden [21] it is not true for all terms of such a category. They may even lead to associative relationship. So associative relationship was also established between terms which, were indented under a term in the same facet, e.g.,

CATALOGUE CODES

RT CATALOGUING

CATALOGUING

RT CATALOGUE CODES

DISPLAY

The use of faceted classification schedule for thesaurus construction helps to display the thesaural entries in two parts i.e.,

a) The systematic part: it is a faceted depth classification schedule.

b) The alphabetical part: this displays the thesaural entries in alphabetical format showing the equivalence, hierarchical and the associative relationships of the descriptors. No descriptors in the thesaurus accompanied class numbers.

An index to the depth classification schedule was prepared. It lists all the isolates and their speciators. This index is an alternative mechanism to link the alphabetical thesaurus to the systematic part, e.g., the descriptor 'catalogue' in the thesaurus will take the searcher to the following entry in the index of the classification schedule.

Catalogue [1MP1] 6c; (Sp1) to [1E] ze; (Sp1) to [2E], ze

From this entry the user can reach to respective concept in the depth classification schedule.

Apart from this function the index helps to find concepts from the depth classification schedule, when the schedule alone is to be used.

EVALUATING OF THE THESAURUS

In order to evaluate the thesaurus, subject headings were prepared for subjects represented in 500 assorted micro documents. The derivation of subject headings was easy and satisfactory. The subject headings derived were expressive enough to disclose the subject of the micro document.

CONCLUSIONS

Faceted depth classification schedules are very helpful in thesaurus construction and particularly help in establishing the thesaural relationships, as discussed below:

1. The isolate terms forming chain divisions in the depth classification schedule mostly represent hierarchical relationship.
2. Establishment of associative relationship has been one of the most intricate aspects of thesaurus construction work. However, use of faceted classification schedule makes this work simple. This is because isolate terms across facet in a classification schedule represent associative relationship. This method ultimately brings consistency in the establishment of associative relationship.
3. The depth classification schedule based on speciators further helps in the establishment of thesaural relationships. This is because the speciators express a pattern. For example, the speciator representing 'types' generate narrower terms to compound term formed by its combination with the isolate.
4. Analogous to the above concept the speciator representing 'cause', 'methods' and 'tools', generate associative relationship.

REFERENCES

1. ROWLEY (J E) and FARROW (J). Organizing knowledge: an introduction to managing access to information. 3rd ed. 2000. Gower; Aldershot.
2. RANGANATHAN (S R). Prolegomena to library classification. 2nd ed. 1957. The Library Association; London
3. RANGANATHAN (S R). Design of depth classification: methodology. *Library Science.* 1; 1964.
4. HULME (F W). Principles of book classification. *Library Association Records.* 1911; 13-14.
5. RANGANATHAN (S R). Prolegomena to library classification. Vol.1. 3rd ed. 1989. Sarada Ranganathan Endowment for Library Science; Bangalore.
6. Ibid
7. RANGANATHAN (S R). Elements of library classification. 3rd ed. 1962. Asia; Bombay.
8. *Ibid.*
9. UNESCO. Guidelines for the establishment and development of monolingual thesauri for information retrieval. 2nd rev. ed. 1981. UNESCO; Paris.
10. BRITISH STANDARDS INSTITUTION. British standards 5723:1979- Guidelines for the establishment and development of monolingual thesauri. 1979. BSI; London.
11. ROY (A K). Thesaurus in information system. In T. N. Rajan. Ed. Indexing systems: concepts, models and techniques. 1981. IASLIC; Calcutta.

12. GUINCHAT (C) and MENOU (M). General introduction to the techniques of information and documentation work. 1983. UNESCO; Paris.

13. ROWLEY (J E). Organizing knowledge: an introduction to information retrieval. 1992. Gower; Aldershot.

14. AITCHISON (J), GILCHRIST (A) and BAWDEN (D).Thesaurus construction and use: a practical manual. 4th ed. 2000. Aslib; London.

15. ROGET (P M). Thesaurus of English words and phrases. 1987. Longman; England.

16. VICKERY (B C). Thesaurus - a new word in documentation. *Journal of Documentation.* 16, 4; 1960.

17. LANCASTER (F W). Vocabulary control for information retrieval. 1972. Information Resources Press; Washington, D.C.

18. LANCASTER (F W). Thesaurus construction and use: a condensed course. 1985. UNESCO; Paris.

19. DAVIS (C H) and RUSH (C L). Guide to information science. 1979. Greenwood; Westport.

20. *op. cit. 9*

21. *op. cit. 14*