SEMINAR ON THESAURUS (1975). Paper AE

CROSS-REFERENCING AND THESAURUS MAINTENANCE

A S RAIZADA, R SATYANARAYANA, M RAMACHANDRAN, and YASH PAL,
INSDOC, New Delhi 110 012

The development of thesauri in science and technology is briefly discussed. The important aspects of the structure and format of a thesaurus are mentioned. The need for revising a thesaurus, the structural changes that may occur in course of time, and the problems of maintaining a thesaurus are indicated. It is the need that should determine the frequency of revision, rather than any other exigency. The types of relations implied in the cross-reference structures in thesauri are briefly discussed. The cross-reference structures in TEST, Thesaurofacet, INIS Thesaurus etc, are examined. Stresses the point that the policy regarding the cross-referencing in thesauri should be stated explicitly, if proper and useful studies are to be undertaken regarding the cross-reference structures, and points out that such a policy statement is of help in the maintenance of thesauri and usable form. The two types of approaches suggested Kochen and others for the study of properties of cross-reference of structures are mentioned.

0 INTRODUCTION

The inadequacy of conventional information systems to meet effectively the challenge posed by the accelerated rate of growth of information in science and technology has given rise to modern high speed retrieval systems . Proper functioning of these systems depends upon the existence of certain basic tools. One such tool is the thesaurus.

1 THESAURUS

For the purposes of this paper, "a thesaurus is a controlled list of terms, with indication of conceptually associated terms, for use in information retrieval systems". (8).

11 Need

Specialists in every field require retrieval of specific parts of documents; the contents of these documents must be concisely symbolized by words which serve as access point to a collection of documents of any size. "Since meanings of words vary according to the indexer or searcher and since individuals associate concepts with different symbols, it seems useful to have at least one constraint in the semantic keleidoscope" (9). The word content of a document is bound to be variable, while a controlled vocabulary (thesaurus) is an invariant, which both the indexer as well as the searcher may use to provide entry to a document or to information stored in a system.

12 Development of Thesauri in Science and Technology

Scientists have always been concerned with precision of terminology in scientific communications. Many thesauri in science and technology have resulted from this concern (9). Some major landmarks are mentioned below.

One of the 'classics' in this area is the Thesaurus of Engineering Terms published in 1964 by the Engineers Joint Council (EJC). The major objective of EJC in formulating the thesaurus was to evolve standard techniques of information handling. A precursor of EJC thesaurus was the Armed Services Technical Information Agency's Thesaurus of ASTIR Descriptors published in 1960 and revised in 1962. A third effort in this direction was made by the American Institute of Chemical Engineers (A I Ch E) in the formulation of Chemical Engineering Thesaurus in 1960. In 1965, a major effort was made to revise the EJC and ASTIA thesauri under the Project LEX and the outcome was TEST - Thesaurus of Engineering and Scientific Terms, published in 1967 jointly by the EJC and the Department of Defence (USA). The Chemical Engineering Thesaurus naturally got absorbed into this effort. A number of micro-thesauri have been compiled using the TEST data base. Some of these are: Thesaurus of Textile

Technology, Paper and Pulp Thesaurus and Thesaurus of Descriptors (Water Resources Development).

With the belief that a large scale thesaurus could be effectively constructed only through the use of classification, particularly facet analysis, the English Electric Co brought out a novel device called thesauro-facet. It is a completely integrated thesaurus and faceted classification scheme, the aspects of which could be used for any of the tasks normally associated with either thesauri or faceted classifications {7). Thesaurofacet covers the entire range of science and technology, though some subject areas are treated in greater depth than others. The main principle behind this thesaurus is that the alphabetical index to the faceted classification is extended and modified so that it becomes a thesaurus in its own right (7).

Another important development in thesaurus construction is the publication of INIS : Thesaurus under the aegis of IAEA in the year 1970. The terms listed in this thesaurus have been derived from the indexing of about 987,000 abstracts in the field of nuclear science and technology, by the staff of EURATOM. Except for minor changes, the terminology of the 1969 edition of the EURATOM Thesaurus has been used in the formulation of this thesaurus (10).

2 STRUCTURE AND FORMAT OF A THESAURUS

A thesaurus not only groups terms by concept or alphabetically, but also establishes and exhibits relationships. The types of relationships in a thesaurus are (a) semantic factors, (b) hierarchical relationships, (c) associate relationships, and (d) syntactical relationships.

21 Structure

All the above mentioned relationships must be encompassed in a suitable conceptual structure and this structure must be displayed in the format. There are two approaches to this problem. One is to have a classified structure such as the one followed in traditional schemes of classification. The other is to adopt a hierarchical pattern with associate relationships. ESC and TEST follow this approach while the Thesaurofacet is designed on the classified approach.

22 Format

The format of a thesaurus is one of the important aspects which enhances its utility in an IR system. The recognised patterns are: the pattern adopted by Roget and the ones followed by TEST and Thesaurofacet. In general, a thesaurus format comprises of the following functional parts: (a) introduction, (b) thesaurus of terms, and (c) indexes.

221 Introduction

The introduction to a thesaurus usually mentions the subject coverage. It may also mention the degree of specificity of the concepts included and the type of thesaurus prepared in relation to the existing thesauri. The description of the conceptual structure, instructions pertaining to its usage, information on the procedures of updating, etc also form an integral part of the introduction (16).

222 Thesaurus of Terms (or the Main Part)

This part lists the terms selected. An entry is made for every term and all the entries thus made are arranged in an alphabetical sequence. The main part also contains information given for each entry, lead-in-terms and descriptors indicating UF, BT, NT and RT relationships.

223 Indexes

The alphabetical index, the subject category index and the hierarchical index etc are arranged in such a way that they provide a multiple approach to the main part of the thesaurus. These are essential for identifying all the relevant descriptors and for locating supplementary information in the main part.

3 FUNCTIONAL MAINTENANCE OF A THESAURUS

31 Need

In live disciplines new concepts and terms arise, some terms become obsolete and concepts themselves change and also their relationships with the existing concepts get modified as the discipline is cultivated further. Even if no such developments take place, it is likely that while indexing a large number of documents, one encounters terms that have not been noticed earlier when the thesaurus was first constructed. To cope with these developments, a thesaurus must be updated on

a continuing basis, otherwise it would become outmoded. Therefore, maintenance of thesaurus in a readily usable form is a very important aspect. Procedures for updating a thesaurus are closely related with the operation of an information system as a whole.

32 Types of Changes

Several types of change may take place in the framework of a thesaurus in course of time. Identification of these changes is necessary for a proper understanding of the problems associated with its maintenance. They are (a) changes in the synonym-homonym structure; (b) changes in the lead-in part of the classificatory structure; and (c) changes in the indexing language (16). These changes may lead to (a) introduction of a new descriptor; (b) elimination of a descriptor; (c) sub-divisi on of an existing descriptor into a number of narrower ones; (d) change in definition and usage of the descriptor, especially leading to a change in delineation between two descriptors, changing the definition of both; (e) addition or elimination of a hierarchical relationship, especially assigning the descriptor to a different group in the hierarchy; and (f) addition or elimination of related term (RT) relationship.

A change of any one of the above types may lead to a chain of changes resulting in a revision of the thesaurus as a whole.

33 Frequency of Revision

Authoritative updated version of a thesaurus has to be brought out at regular intervals. Delay in revision poses problems to information centres concentrating on a rapidly developing discipline and using a thesaurus for information handling. One approach to this problem is for the information centre using a particular thesaurus to develop modifications in its specialized field with the concurrence of the body responsible for updating the thesaurus at short intervals, These modifications are incorporated into the thesaurus when a subsequent.edition is brought out by the agency concerned. It is the need that should dictate the frequency rather than any other exigency.

CROSS-REFERENCES

A thesaurus establishes and exhibits relationships between terms listed in it. Let us examine what these relationships are and by what method or methods they are exhibited in a thesaurus structure.

41 Types of Relationship

The main types of relationship that we come across in a thesaurus arc : (a) morphological relationships; (b) synonyms; (c) antonyms; (d) preferred terms; (e) inclusion relations; and (f) others. A proper understanding of these relationships is essential (13) to know the cross-reference structure followed in a thesaurus.

411 Morphological Relationships

Spelling variations in the terms; For example, paediatrics_ pediatrics; colour - color; disc - disk, belong to this category. Such terms are identical in meaning and are therefore strongly linked. This kind of relationship has to be exhibited by means of a cross-reference. In a manually produced thesaurus, cross-referencing of word variants does not pose a serious problem.

Permutations and combinations of words such as Central Nervous System - Nervous System, Central; Radar Antennas - Antennas, Radar, etc, also belong to this group. Abbreviated expressions are accepted in a thesaurus language whenever the possibility of ambiguity is small. The See (USE) reference has the function of leading the searcher from the abbreviated expression to the complete one and vice-versa. This is necessary because of the fact that the two terms cannot be used interchangeably as index terms.

412 Synonym s

In scientific and technical work, different words or expressions may have the same meaning. For example, in chemical jargon, common salt is sodium chloride and aspirin is acetyl salicilic acid. The problem is the choice of a preferred term (PT) for use in a thesaurus for indexing purposes in IR systems. The relationship between the synonymous terms has to be indicated in a thesaurus by means of appropriate cross-referencing, such as, by using the abbreviation ST for Synonymous term.

413 Antonyms

Antonyms find *place in* cross-reference structures in thesauri. Though, in general practice, antonyms convey opposite meanings, in fact they form two opposite ends of a conceptual continuum (13); for example, the terms 'expansion' and 'contraction'. This relationship between antonymous terms is indicated in a

thesaurus using the abbreviation RT (Related Term).

414 Preferred Term

If it is assumed that two given synonymous terms or antonymous terms linked by symmetrical relation can be used interchangeably to index documents, the problem is to choose one of these terms as a preferred term (PT). Though, by convention, the most used term is chosen as a preferred term and a cross-reference is provided to and from the less used one, scientific studies to establish the validity of this practice are not available (13). This type of relationship is indicated in a thesaurus by UF. For example, Fungus OF Beech indicates that between the two terms 'fungus' and 'beech', the former is the preferred term. Another entry, Beech USE Fungus.

415 Inclusion Relations

Four different types of inclusion relations are found in cross-reference structures (13). The first one corresponds to the concept of class inclusion or species-genus relationship. For instance, the class Nuclear Scientists is contained in the class Scientists. The second type of relation pertains to individuals as members of a class. For example, Dr H J Bhabha is a member of the class Nuclear Scientists. Dr Bhabha is a scientist but, all scientists need not be Nuclear Scientists. Thus, a characteristic of class inclusion not but of class membership is that classes can be arranged hierarchically in such a way that a class is generic with respect to the class at a lower level and specific with respect to the terms at a higher level. This sort of relationship is adopted in a thesaurus following a classificatory structure such as 'thesaurofacet'.

The third type of inclusion relation pertains to 'part to whole' relationship. Cross-references of this kind are based on structural-spatial relation.

The fourth type is topic inclusion relation. It concerns the relationships between two aspects of knowledge one inclusive of the other. For example, thermodynamics and physics.

These four types of relations occur in varying degrees in cross-reference structures adopted in thesauri. Cross-references of inclusion type provide converging links from various terms located at a lower hierarchical level (NT) towards a single term (BT) situated at a higher level. This type of cross-reference enables the search procedure aiding the user of thesaurus to avoid multiple specific searches with a single search under a broader term (BT).

416 Other Relationships

The foregoing enumeration covers only the basic relationships recognised between terms and their cross-reference structures. There can be other types of relationships such as cause and effect, product and producer, etc, At times, it is rather difficult to define the existing relationship between concepts. Nevertheless, we must be aware of such relationships and indicate them by suitable kind of cross-references.

5 CROSS-REFERENCING PATTERN IN EXISTING THESAURI

In the construction ESC, TEST, Thesaurus facet, INIS Thesaurus, etc. the above mentioned relationships have been followed in varying degrees. A close study of the above thesauri reveals that only two or three types of relationships have been primarily taken into consideration in providing cross-reference structures. These are (1) See type (USE); (2) Specific-Generic type (NT, BT), and (3) Related Term (RT.) type.

Each of the above named thesaurus has provided guidelines explaining the nature of the cross-referencing pattern they have adopted. Some examples illustrating the network of cross-references in TEST, Thesaurus of Pulp and Paper Terms, and INIS Thesaurus, are given in Tables 1 and 2 (1),

511 Observations; TEST and INIS Thesaurus

For a given descriptor, the levels of cross-referencing are different; for example, for the term at SN 1, TEST has provided 6 related terms. If the particular conceptual relationship between (1) and the related terms is to be presented, a cross reference is necessary in each case. The hierarchy of these terms with the descriptor is not explicit. In the case of INIS, all the 6 terms are narrower than the descriptor indicating the generic to specific relationship. This establishes that INIS Thesaurus gives emphasis to specificity of the concepts. At SN 4, TEST gives the descriptor 'Erythrocytes' as a related term to 'Hemagglutination', while the INIS Thesaurus shows the same as a broader term with respect to the descriptor. TEST lists 'Agglutination' as a

51 Table 1: Cross-Reference Patterns: TEST and INIS Thesaurus

	Descriptor	TEST	INIS
1	Excitation	Excitation RT Activation Actuation Electron transition Emission Nuclear capture Relaxation Time	Excitation NT Activation Energy Coulomb Excitation De-excitation Nuclear Temperature Optical Pumping Stimulated emission
2	Moderators	Moderators OF Nuclear Reactor Moderators RT Beryllium Graphite Heavy Water Nuclear Reactor Materials Thermal Column	Moderators NT Moderator Fuel Ratio Reactor Materials Sigma Piles Thermal Column
3	Fourier Transformations	Fourier Transformations BT Analysis (Mathematics) Functional Analysis Functions (Mathematics) Integral Transformations RT Fourier integrals	Fourier Transformations BT Integral Transformations Integrals Mathematics NT Inverse Fourier Transformation
4	Hemagglutination	Hemagglutination BT Agglutination RT Erythrocytes	Hemagglutination BT Erythrocytes Immunity
5	Thermonuclear Reactions	Thermonuclear Reactions BT Nuclear Reactions RT Nuclear Fusion -- Pinch effect -- Plasmas (Physics) -- Stellorators -- Thermonuclear energy	Thermonuclear Reactions NT Gravitational collapse -- Gravitational Radiation -- Project Sherwood Thermonuclear Explosions
6	Thermistors	Thermistors BT Resistors Semiconductor Devices NTInfra-red Thermistors RT Fixed Resistors-Variable Resistors variators	Thermistors BT Resistors Temperature
7	Nuclear Emulsions	Nuclear Emulsions BT Dispersions Emulsions Photographic Emulsions Photographic Materials RT Radiation counters Radiation Measuring Instruments	Nuclear Emulsions NT Agfa Emulsions Herschel Effect Loaded Nuclear Emulsions Nikfi Emulsions
8	Nuclear Induction	Nuclear Induction BT Nuclear Properties RT Nuclear Magnetic Resonance	Nuclear Induction BT Induction Magnetic Moments Nuclear Magnetic Resonance Nuclei
9	Neutron flux	Neutron flux BT Flux (rate) Particle flux Rates (per time) RT Neutron flux density Neutron irradiation Radiation shielding	Neutron flux NT Adjoint flux Flux tilting Neutron leakage Disadvantage Factor

52 Table 2: Cross-Reference Patterns: TEST and Pulp and Paper Thesaurus

	Descriptor	TEST	Pulp and Paper Thesaurus
1	Abrasion Resistance	Abrasion Resistance BT Mechanical Properties Wear Resistance RT Abrasion Resistance coatings Abrasion Resistance Steels -- Hardness	Abrasion Resistance OF Abrasion Loss Wear Resistance NT Oil-Rub Resistance Scuff Resistance Wet-rub Resistance BT Mechanical Properties RT Abrasion Erasing Quality Mechanical Tests Rubbing Wear Tests
2	Brighteners	Brighdeners NT Optical brighteners RT Additives Bleaching agents Dyes	Brighdeners OF Whiteners NT Optical Brighteners RT Additives Agents Colours Dyes Fillers Fluorescent Dyes Modifiers Pigment Taint
3	Clarification	Clarification UF Dehazing BT Separation RT Clearing Cleaning Coagulation Effluents Sedimentation Skimming Straining Thickening Wader treatments	Clarification OF Dehazing BT Sepa rat ion RT Clearing Cleaning Skimming Thickening
4	Density Measurement	Density Measurement RT Aerometers - Density (Mass & Vol) Hydrometers Weight measurement	Density Measurement BT Measurement RT Chemical analysis Density Density meters Dimensional measurements Materials testing Weight measurement
5	Fibre Boards	Fibre Boards NT Box Board Container Board Paper Boards Press Boards Wall Board RT Building Boards Paper products Particle Boards Wood Products	Fibre Boards NT Hard Boards Soft Boards BT Building Boards RT Fibre Board Drums Insulating Boards Insulation Masonite Paper Boards Thermal Insulation Vulcanized Boards Wall Boards
6	Hemicelluloses	Hemicelluloses BT Carbohydrates Polysacchorites RT Cellulose	Hemicelluloses BT Natural Polymers Polymers RT Aragon Beta cellulose Carbohydrates Caustic solubles Cellulosons Cellulose Gamo cellulose Hexosans Xylan

concept to 'Hemagglutination'. This indicates that cross-referencing level with respect to the term 'Erythrocytes' is differently conceived in the two thesauri.

In the same way, if we examine the other descriptors listed in the table, certain differences in the conceptual framework of the terminology leading to different levels of cross-referencing can be noticed_

521 Observations : TEST and Pulp and Paper Thesaurus

If we consider the descriptor 'Brighteners' (SN2), TEST provides one narrower term 'Optical brighteners' and three related terms. In the Pulp and Paper Thesaurus, a synonym 'Whiteners' is provided as leading term while 'Brighteners' has been indicated as a preferred term. In both the thesauri, the term 'Optical brighteners' has been conceptually linked as a narrower term. So far as related terms are concerned, two of the terms provided by TEST figure in Paper and Pulp Thesaurus as well. The latter provides seven more related terms. But the hierarchical relationships of these related terms with the descriptor 'Brighteners' is not made explicit. From this we may infer that the Pulp and Paper Thesaurus being a specialised one, has provided various types of relationships that a description could have indicated its specialized usage.

Again, if we consider the descriptor 'Density measurement', the treatment of the related terms is somewhat deeper in the Pulp and Paper Thesaurus. In both the cases, the terms 'Density' and 'Weight measurement' figure as related terms. There is no similarity among the broader terms provided.

A study of all the descriptors provided in the table gives an idea regarding the levels of cross-referencing provided in the two thesauri.

6 CONCLUSIONS

Some information scientists are of the opinion that provision of more cross-references would make a thesaurus a more effective tool in IR systems. This appears to be reasonable on an a priori basis as cross-referencing increases the number of entry points in a system. But proper studies to support this contention are yet to be made. However, cross-references are helpful means by which effective links can be maintained between the changes in technical terminology and thus aid in the maintenance of thesauri in usable form on a continuing basis. Though the number of entries increases due to the cross-referencing pattern adopted, the problems posed by such a contingency are better met with larger budgets. We feel that it is not possible to explicitly exhibit the relationships between the terms included in a thesaurus without taking recourse to cross-referencing.

Two types of approaches have been suggested by Kochen and others (13) for the study of properties of cross-reference structures. The first consists of looking at cross-reference structure as a directed graph, in which the nodes are index terms and the links are relations between index terms. The number of nodes and links, the distribution of single versus multiple links, the ratio of direct to indirect pointers may then be the basis for comparing different cross-reference structures. Connectedness and accessibility are two measures for defining the 'level of cross-referencing' of a thesaurus.

The second approach consists of considering the linguistic aspects of cross-reference structures. In this case, it is essential to examine the types of relationship which link the terms of cross-reference, the strength of the bond between the two terms of a particular relation on the basis of common properties. It is also necessary to establish principles for determining the choice of preferred terms and study about the relations utilized by the users of a thesaurus as compared to the relations embodied in the cross-reference structure. Studies conducted on this basis should determine the future course of research in the field of thesaurus. These studies can be more effective if the policy of cross-referencing is explicitly stated at the time of construction of a thesaurus. Such a policy statement helps in the updating work of the thesaurus. In our study of the above mentioned thesauri, we were not able to find out this statement and the scheme of cross-referencing had to be inferred from the pattern of the thesauri. If the cross-referencing policy is explicitly known, then we will be in a position to determine the optimal cross-reference structure for a particular thesaurus to be used in a particular information system.

7 ACKNOWLEDGEMENT

The authors express their gratitude to Shri S Parthasarathy, Scientist-in-charge, INSDOC, for providing necessary encouragement in the preparation of this paper.

8 BIBLIOGRAPHICAL REFERENCES

1 AITCHISON (K) and GILCHRIST (A). Thesaurus construction: A practical manual. 1972.

2 AMERICAN INSTITUTE OF CHEMICAL ENGINEERS. Chemical engineering thesaurus. 1961.

3 BLAGDEN (S). Structured thesaurus. (Aslib Proc. 23; 1970; 139-43).

4 CAMBELL (D J). EJC thesaurus of engineering terms. (J Doc. 21; 1965; 136-9).

5 ESC. Thesaurus of engineering terms. 1964.

6 ESC and US DEFENCE (Dept of -). Thesaurus of engineering and scientific terms. 1967.

7 ENGLISH EIECCENTRIC CO. Thesaurofacet : a thesaurus and faceted classification for engineering and related subjects. 1969.

8 FID NEWS BULL. 20; 1970; No. 106.

9 GILCHRIST (A). Thesaurus in retrieval. 1971.

10 INTERNATIONAL ATOMIC ENERGY AGENCY. INIS; Thesaurus. 1970.

11 JANE WEINSTEN (S). Biological dictionary preparation, control, and maintenance. (Am Doc. 17; 1966; 190-8).

12 SONES (K F). Basic structures for thesaural systems. (Aslib Proc. 23; 1971; 577-90).

13 COTTON (M) and TAGLIACOZZO (R). Study of cross-referencing. (J Doc. 24; 1968; 173-91).

14 PAPER AND PULP RESEARCH INSTITUTE. Thesaurus of paper and pulp terms. 1965.

15 ROLLING (L). Graphic display devices in thesaurus construction. (Aslib Proc. 23; 1971; 591-4).

16 SOERGEL (D). Indexing languages and thesauri: construction and maintenance. 1974,

17 UNITED STATES. INTERIOR BUREAU OF RECLAMATION. (Department of -), Thesaurus of descriptors. 1963.