SEMINAR ON THESAURUS (1975). Paper CA

THESAURUS - AN IDEAL TOOL FOR VOCABULARY CONTROL IN POST-COORDINATE SYSTEMS : INIS THESAURUS - A CASE STUDY

SK KUMAR, MK RAGHAVENDRA RAO, and VA KAMATH,
Library and Information Services, Bhabha Atomic Research Centre, Bombay 400 085

The complex conceptual relationships contained in multi-dimensional documents present the problem of inconsistency in subject analysis and vocabulary control for efficient retrieval is stressed. The development of Thesaurus, an important contribution for effective vocabulary control in mechanised IR systems which is among the many developmental efforts in this direction, is highlighted. The development, structure and design of INIS Thesaurus is given in detail. Adaptability of the INIS Thesaurus to a computer-based information handling system is examined

1 INFORMATION ANALYSIS - THE PROBLEM OF CONSISTENCY

Recorded information is becoming increasingly more complex and multi-dimensional in character, It usually involves a complex interrelationship between various substances, devices, procedures and their respective characteristics. Even the simplest records are multi-dimensional and it is usually difficult to predict which point of view (or dimension) will be pertinent in a particular search. In order to organise the information contained in graphic records in such a way as to facilitate its identification on demand, it is necessary to analyse the multi-dimensional records in a consistent manner. Common experience suggests and experiments confirm that a person does not always make the same choice when faced with the same option, even when the circumstances of choice seem in all relevant respects to be the same. This statement summarises one of the basic problems of conducting consistent content analysis (indexing, classification etc), which serve the purpose of an information retrieval system. It is not possible for a single analyst, much less a team of analysts, to exercise absolute consistency in the way in which they conduct analysis of graphic records. Although the principles of analysis are basically the same, the techniques adopted and the particular retrieval device used may help in saving money or time or aid in effective searching. This implies that no analytical method can be considered as 'perfect'. However, some means have been experimented for overcoming the effects of inconsistency, either through the use of redundancy in analysis, vocabulary control, coding methods, cross reanalysis, using alternative search strategies etc. These subject analysis and control methods tend to improve considerably the indexing consistency.

2 PRE-COORDINATE VS POST-COORDINATE SYSTEMS

The subject matter of any document tends to be complex and therefore the vocabulary of any modern IR system should be capable of expressing such complexity. The first attempt in this direction was through pre-coordination systems such as Classified Card Catalogues, Alphabetical Subject Catalogues, etc However, these systems failed to display sufficiently the conceptual relationships of the complex subject matter of documents, particulary those which contained multi-disciplinary subjects. Consequently, the pre-coordinate system could neither provide the required flexibility nor were suitable for multi-dimensional searches. This led to the development of post-coordinate systems such as Subject Heading List, Authority List, Controlled Vocabulary, Thesaurus, etc. The pioneering contribution in the development of postcoordinate system was made by Batten, Moores and Taube. They were responsible for the development of Optical Coincidence System, Descriptor System, and Uniterm System respectively. In all these systems, careful vocabulary control was riot given much importance. Later, it was realised that the use of single word systems based on uncontrolled language would present certain problems such as : synonyms, homonyms, generic search, syntactical relationship, viewpoint, etc. Gradually, it was i-ecogmseuYh'at'post-coordihate system would benefit a great deal if proper vocabulary control devices were used and thus the THESAURUS was born.

3 THESAURUS - DEFINITION AND STRUCTURE

Thesaurus is a terminological control device used in translating from the natural language of documents, indexers or users into a more constrained 'System Language' (document language, information language). In other words, 'Thesaurus' is a controlled and dynamic vocabulary of semantically and generically related terms which cover a specific domain of knowledge. This is the definition accepted by the UNESCO. Thesaurus which has been widely adopted for vocabulary control in modern post-coordinate systems, lists descriptors alphabetically, endeavours to control synonyms and homographs and displays generic, specific and other relationships between terms. In fact, the Thesaurus is very similar to the List of Subject Headings except that the device used for displaying the relationship between terms are somewhat different. This is because the application; of these tools vary from one another. The Subject Heading List is generally used as an indexing tool employing manual information retrieval systems whereas the structure of the Thesaurus is specially designed for use in mechanised IR systems.

4 DEVELOPMENT OF THESAURUS

According to Joyce and Needham, investigators at the Cambridge Language Research Unit in England began to discuss the applicability of the Thesaurus concept to information retrieval in 1956. In this context, the word appears to have been used first in print by Hans Peter Luhn of IBM in 1959- According to F W Lancaster, the first Thesaurus actually used for controlling the vocabulary of an information retrieval system was developed by the Dupont, USA, and appears to date from about 1959. The first widely available thesauri were the Thesaurus of ASTIA Descriptors (Department of Defence, 1959) and the Chemical Engineering Thesaurus (American Institute of Chemical Engineers, 1961).

The first edition of Medical Subject Headings appeared in I960, but the second edition, 1963, was the first designed specifically for use in the post-coordinate machine-based system, MEDLARS. The influential Thesaurus of Engineering Terms of the Engineers Joint Council (EJC) was published in 1964. In the decade I960 - 1970, many other thesauri were developed, some widely available while others solely for internal use within the organisation.

5 INIS THESAURUS

International Nuclear Information System (INIS) is a computer-based, mission-one ted, decentralised information handling system in the field of nuclear science and technology sponsored by the International Atomic Energy Agency (IAEA) operating in Vienna since May 1970. INIS was created in a spirit of international cooperation in which some 58 countri and international and regional organisations a participating. This is the first international information system operating in a highly decentralised set-up with regard to input prepara tion. Among the many Reference Manuals made available to the inputting centres for processing the input data, the INIS Thesaurus i« one. This thesaurus represents a homogene logical system equally useful for subject indexing and for machine search and retrieval. It has been developed over a period of years durir which many technical thesauri, descriptor lis' and subject heading lists were consulted. IMS Thesaurus covers not only nuclear physics and reactor technology, which it covers in de but also related topics such as isotope technology, fabrication and use of nuclear materials and instruments, radiochemistry and radiobiology, to a lesser degree. At present, there are 14, 123 accepted terms (descriptors) and 4, 013 forbidden terms (non-descriptors) in the INIS Thesaurus. The terminology in the IMS Thesaurus is listed alphabetically but with each entry, the full 'Word Block' is displayed, giving all the terms associated with that particular entry, (Appendix 1 ; a typical page from the INIS Thesaurus shows different kinds of inter-relationships).

6 STRUCTURE OF THE INIS THESAURUS

Thesaurus is one of the most important constituent of an information system. It is not only an important tool in indexing, but also an essential key for retrieval. Thesaurus is like a living organism which, to keep up with its tasks, must continuously renew itself with the help of a computer. In other words, the thesaurus must be structured in such a way that introduction, deletion or substitution of terms can be easily carried out at any time without any loss of the Information stored in the computer memory. The most important function of the thesaurus is to serve as a tool in information retrieval. Descriptors chosen from the thesaurus must give clear indication of the information and data content of the document. To do this, their meaning must be well defined and completely unambiguous. This semantic definition must be provided in the thesaurus by means of a structure which is given to the terminology. The inter-relationship between individual descriptors should be brought out. There are three types of inter-relationships: Preferential, hierarchical, and affinitive. Preferential indicators (See, Use, Used for = UF, Seen For = SF type cross references) identify the preferred choices in cases of semantic ambiguity. In other words, they are used when the meaning of descriptors overlap substantially. These cross references are useful for referring a forbidden term to a descriptor(s).

Hierarchical indicators (Broader Term = BT, Narrower Term = NT) identify the semantic relationship existing between descriptors of different specificity levels in the same hierarchy of concepts.

Affinitive indicators (Related Term(s) = RT) are employed to refer from one descriptor to others that are related in concept but are neither consistently, hierarchically nor preferentially related. (See Appendix 1 for examples).

The purpose of displaying descriptors showing inter-relationships is mainly fourfold :

a) The forbidden, broader, narrower and related terms associated with each descriptor semantically define the meaning in which the descriptor is to be used to represent a concept in a piece of literature.

b) The narrower terms may suggest more specific, appropriate terms than the one being considered.

c} The INIS computer programmes are such that an automatic procedure called 'upposting' assigns to the piece of literature all broader terms associated with the particular descriptor chosen.

d) The related terms provide the essential connection to semantically, functionally or otherwise related concepts to form a network of pathways along which searches for appropriate terminology can be made

To an indexer, it means that, when a descriptor suggests itself as appropriate, his study of the work block for that term will tell him whether the intended term is in its proper meaning ; he could be led to consider more specific descriptors which may be appropriate ; he will be aware of all the broader terms which will be automatically assigned as a consequence of his actually using the descriptor being considered and he will be made aware of the related concepts which it may be useful to consider as appropriate.

It may appear peculiar to have 'Forbidden Terms' in the thesaurus. However, forbidden terms form an integral part of the thesaurus in its purpose of exercising language control by defining choices between words used in natural language Forbidden terms lead the indexer from the unacceptable form of the word to the accepted one for expressing the same concept.

7 GENERIC POSTING

INIS is computerised expecially with respect to the thesaurus. Descriptors and structures are stored on magnetic media Thus, the reference structure of the INIS Thesaurus is part of the computer software. The computer is used to perform what is known as 'Generic Posting'. This generic posting, that is, transfer of information to higher generic levels, provides additional access points for retrieval The enormous advantage of this posting on to generic and cumulative terms obviates the need to include long lists of alternative terms in query formulation. A series of automatic checking and correcting programs does the following functions: (a) 'USE' operators (forbidden terms) are replaced automatically by the corresponding authorised terms, (b) term with 'SEE' operators (related terms) are printed out for manual checking, (c) terms with 'ADD' operators are automatically posted to higher generic levels, (d) indexed terms occurring second time in an analysis are automatically eliminated, (e) minor misspellings of indexed terms are automatically corrected, (f) terms of higher hierarchical levels than specific descriptors are automatically assigned (broader terms) for use in retrieval exclusively.

8 AID TO INDEXERS

The structure of the thesaurus may he excellent, but if the indexer or retriever can-not make correct use of it, then it will be of little benefit to anyone. The indexer must acquire the necessary knowhow. For this purpose, indexing rules have been drawn up in the INIS ; Manual for Indexing (9). For high quality indexing, it is a pre-requisite that indexers have good subject knowledge. In order to achieve high consistency, the controlled vocabulary (thesaurus) from which descriptors are to be selected needs an understandable and predictable structure and clear presentation. The easier it is for an indexer to be led to the appropriate terminology in the thesaurus, the more likely he will produce consistent indexing. Therefore, the criteria on which the thesaurus is constructed should be unambiguous and clear to all indexers. For this purpose, a special document entitled Guidelines for the Development and Maintenance of the INIS Thesaurus (10) is at the disposal of the indexers.

91 APPENDIX - 1

Reproduction of a typical page from the INIS Thesaurus

RADIATION INJURIES [1,121; 1,182}
(For damage to molecules of biological significance we CHEMICAL RADIATION EFFECTS or STRAND BREAKS.)
UF -radiation damage (biological)
UF+ -delayed radiation injuries
UF+ -early radiation *injuries*
BT1 **biological radiation effects**
BT2 **biological effects**
B12 **radiation effects**
BT1 Injuries
BT2 **diseases**
NTI **osteoradionecrosis**
NT1 **radiation burns**
NTI radiodermatitis
RT **biological indicators**
RT **biological repair**
RT host-cell reactivation
RT photoreactivaiton
RT **radiation syndrome**
RT **radiobiology**

RADIATION MONITORING [1,527] 1,916]
UF control (radioactivity)
UF monitoring (radiation)
UF -surveillance (radioactivity)
UF survey (radktactirity)
BT1 monitoring
NTI **aerial monitoring**
NTI **personnel monitoring**
RT **aerosol monitoring**
RT **alarm systems**
fir dosemeters
RT **dosimetry**
RT **exposure ratemeters**
RT **inspection**
RT **radiation detection**
RT **radiation protection**
RT **radioactivity**

Preferential relationship

Hierarchical relationship

Affinitive relationship

RADIATION MONITORS [170; 434)
UF alarm dosemeten
RTI measuring instruments
NTI exposurer ratemeters
NTt **liquid *a,* illumination monitors**
NT! **neutron monitors**
Nil **surface contamination monitors**
NTI **survey monitors**
RT **alarm systems**
R7' **dosemeters**
RT **radiation detectors**
RT" **radioactivity**

RADIATION PRESSURE [68; 68]
UF -pressure *(radiation)*
RT electromagnetic radiation
RT solar wind

RADIATION PROTECTION (2.075; 2,075)
UF -hearth physics
UF -nuclear safety
UF -protection (radiation)
UF radiation hygiene
UF -radiation safety
UF -radiological protection
UF safety (nuclear)
RT accidents

UF = USED FOR ; SF = SEEN FOR ;
RT = RELATED TERM ; BT = BROADER TERM;
NT = NARROWER TERM.

Figures in square brackets indicate frequency of usage of the descriptor.

92 BIBLIOGRAPHICAL REFERENCES

1 HOLM (BE) and RASMUSSEN (L E). Development of a technical thesaurus. (Am Doc. 12; 1961; 184-90)

2 JOYCE (T) and NEEDHAM {R M). The thesaurus approach to information retrieval. (Am Doc 9; 1958; 192-97)

3 LANCASTER (F W). Vocabulary control for information retrieval, 1972

4 RAGHAVENDRA RAO (M K) and KAMATH (VA). INIS Thesaurus - An ideal model for a computer-based information handling system. (Paper submitted for Publication in Ranganathan Memorial Volume)

5 -- -- INIS - A successful experiment in operating a decentralised, computer-based mission-oriented information system (Presented at the Third International Study Conference on Classification Research (Bombay) (1975)).

6 -- Lectures on documentation, information science, reprography and translation work, 1971

ROSENBERG {K C) and BLOCHEN (C L M). Coinparison of the relevance of key-word-in-context versus descriptor indexing terms. (Am Doc. 19 ; I968; 27-9)

IAEA-INIS-13 (Rev 9)- INIS: Thesaurus

9 IAEA-INIS-12. INIS: Manual for indexing

10 IAEA-142. Guidelines for the developroent and maintenance of the INIS Thesaurus .