AUTOMATED DOCUMENT RETRIEVAL

Lectures for the International Seminar on

DATA BANKS
INFORMATION RETRIEVAL

organized by FAST - Federation of Scientific
and Technical Associations. Milan,8th-27th October 1973.

by Dr. Carlo VERNIMB
Commission of the European Communities
DG.III, Specialized Department INFORMATION SYSTEMS
Luxembourg,29, rue Aldringen

CONTENT

I. THESAURUS CONSTRUCTION AND MAINTENANCE

A. Introduction

The task of documentation is to indicate to the individual scientist or technician from a yearly lot of millions of publications just those few dozens he will be in need of.

For achieving this, the traditional searches from registers containing isolated searching terms will no longer be sufficient. In this day and age their place has largely been taken over by concept coordination (in indexing as well as retrieval). This is true already of systems using peek-a-boo cards. But the real breakthrough of "coordinate indexing and retrieval" is being made today only by means of computers.

Most of the documentation systems working with "coordinate indexing and retrieval" are making use of a Thesaurus. It will be explained why a Thesaurus is used, how it is constructed and maintained, and which effect it will have on retrieval. The INIS Thesaurus, which was developed from the Euratom Thesaurus, will be quoted by way of example; it was developed in particular for a computer-based system, its utility for that system has so far been proved and it is in accordance with international standards for Thesauri.

There are many definitions of Thesauri. The following one has been adopted by UNESCO:

"A thesaurus is a controlled and dynamic vocabulary of semantically and generically related terms which comprehensively covers a specific domain of knowledge. This vocabulary is a systematical and/or alphabetical collection of descriptors, non-descriptors as well as indicators of their relationships."

In principle the same definition but referring more to practice is given by Rolling (1):

A Thesaurus can be defined as a structured vocabulary for use in information storage and retrieval systems.

Three parts of this definition need further elaboration:

1. A vocabulary is a collection of terms.
2. The structure of a vocabulary can be described as a set of relationships between terms.
3. Utilization of a thesaurus in an information system involves a set of rules which take into account the characteristics of the system.

There are three types of thesauri according to the type of terms they consist of:

UNITERMS	UNICONCEPTS	SUBJECT HEADINGS
ARC	ARC WELDING	ARGON-ARC PROCESS
DROP	DROP FORGING	DROP FORGING EQUIPMENT
DELTA	DELTA WINGS	DELTA WING DRAG
FLOW	FLOW RATE	HIGH-TEMPERATURE FLOW RATE DET.N.
PETROLEUM	PETROLEUM	PETROLEUM
THERMOCOUPLES	THERMOCOUPLES	THERMOCOUPLES
ELECTROPHORESIS	ELECTROPHORESIS	ELECTROPHORESIS

Table 1. Types of Thesauri according to the type of Terms they consist of.

B. Usefulness of the Thesaurus

Before the guidelines for the establishment of a Thesaurus are discussed the reasons will be considered why at all a Thesaurus is useful. For the indexer it would be much easier to select terms by underlining appropriate terms in the text of the document or by assigning terms of his own choice instead of searching for permitted terms in a Thesaurus.

According to Vickery (2) the basic reason for applying a

Thesaurus is the immense variety of natural language. Suppose that an index user desires to find documents dealing with oscillation motion, movement to and from, and searches the index for entries under Oscillation. Authors may have discussed this concept under many other names: Vibration, Undulation, Pulsation, Nutation, Swing, Beat, Rolling, Pitching, and so on. If each index entry uses each author's own word, and no links are made between them, the searcher will miss all these entries.

In other words: it is mainly the problem of synonyms which asks for the use of a Thesaurus. This is demonstrated by another example: For searching publications it will not make any difference whether the term used is "plant" or "botany": every scientific publication on plants will be part of botany, and every botanical publication will deal with plants. Thus the words "plant" and "botany" are synonyms for the requirements of documentation. There is no sense to admitting both of them as descriptors. One of them must be established and, later, used. It is evident that the purpose of the Thesaurus is not to facilitate indexing but retrieval.

C. Thesaurus Structure

The rules for the establishment of a Thesaurus, which will be discussed here, will be based on the "Guidelines for the Development and Maintenance of the INIS Thesaurus" which on its part very closely follows the "UNESCO: Guidelines for the Establishment and Development of Monolingual and Technical Thesauri for Information Retrieval" (3).

The most important function of a thesaurus is to serve as a tool in information retrieval. Consequently descriptors must give clear indications of the information and data content of a 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 the structure which is given to the terminology. It becomes imperative therefore, that the interrelationship between individual descriptors be brought clearly into evidence.

These interrelationships are of three types: preferential, hierarchical, affinitive. All three have the property of reciprocity, i.e. when two or more descriptors are related in any way, reciprocal entries should be provided.

Cross references

Interrelationships among terms will be shown by cross references which are grouped according to the three types:

Cross References	Symbol	Type
USE USED FOR SEE... OR SEE FOR	USE UF SEE OR SF	Preferential
BROADER TERM NARROWER TERM	BT NT	Hierarchical
RELATED TERM	RT	Affinitive

1. Preferential relationships

These cross references are employed to refer from a forbidden term to a descriptor (s) and vice versa. They are used when the meaning of descriptors overlap substantially; where different spellings of the same word exist; for synonyms, antonyms and homonyms and, in general, whereever a choice has been made between a number of descriptors, all of which are included in the thesaurus display.

a) USE reference

The USE reference is intended to lead users of the thesaurus from a forbidden term to a descriptor(s) which must be used in its stead. This "Exclusive" reference is employed in a variety of situations:

i) to indicate a preferred synonym

e.g. -columbium

USE NIOBIUM

Here and in the following the forbidden terms are preceded by a minus sign.

ii) to indicate a preference between spelling variations
e.g. -sulphur
USE SULFUR

iii) to expand or explain abbreviations
e.g. -esr
USE ELECTRON SPIN RESONANCE

iv) to reduce the number of precoordinated descriptors by prescribing the use of two or more descriptors to express a concept
e.g. -storage tubes
USE ELECTRON TUBES
AND RECORDING SYSTEMS

v) to express concepts which can be considered synonyms for purposes of indexing and retrieval
e.g. -maxwell-boltzmann distribution
USE BOLTZMANN STATISTICS

vi) to bring together different viewpoints of a conceptual continuum
e.g. -softness
USE HARDNESS

vii) to reflect current terminology, eliminate jargon or exercise a choice for semantic or other reasons
e.g. -electric condensers
USE CAPACITORS

-cosmos
USE COSMIC SPACE

viii) to refer the variation in proper name terminology to one choice
e.g. -einstein diffusion mobility
-einstein ratio
-einstein relation
USE EINSTEIN DIFFUSION RELATION

b) USED FOR REFERENCE

The USED FOR reference (UF) is the mandatory reciprocal of the USE reference and accompanies the descriptor to which the USE reference refers