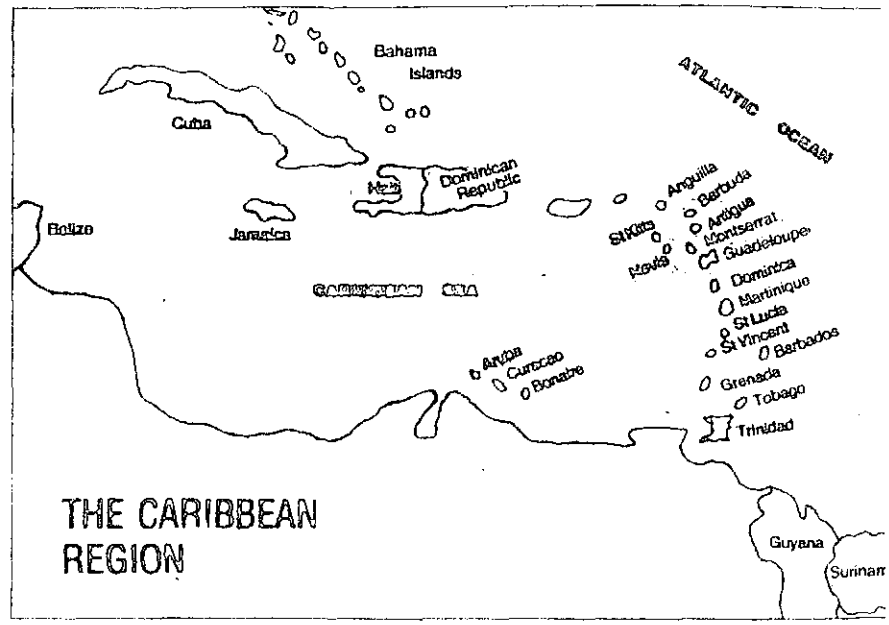
CARIBBEAN DEVELOPMENT AND CO-OPERATION COMMITTEE



THE CARIBBEAN REGION

Distr.
LIMITED

LC/CAR/L.129 (SEM.1/3)
CDCC/CARSTIN/84/3

3 December 1984

ORIGINAL: ENGLISH

ECONOMIC COMMISSION FOR LATIN AMERICA AND THE CARIBBEAN
Subregional Headquarters for the Caribbean

CARIBBEAN DEVELOPMENT AND CO-OPERATION COMMITTEE
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION

CARSTIN Training Workshop/Seminar on
Network Development in the Caribbean
Port of Spain, Trinidad
3-14 December 1984

INDEXING AND INFORMATION RETRIEVAL

Prepared by
Fay Durrant
Consultant

UNITED NATIONS

ECONOMIC COMMISSION FOR LATIN AMERICA Office for the Caribbean

INDEXING
AND
INFORMATION RETRIEVAL

The growth in the production of documentary information has resulted in a need for the organization of material held by an information system. This would enable the user to acquire information at minimum cost of time and money, without having to examine material which is not directly relevant to his information needs.

Information retrieval systems should therefore be seen as a form of communication between the author, or producers of information, and the receivers or users through the channel of the information system.

Indexing is the main component of the encoding process which forms the second stage of the modified version of the Shannon and Weaver [1] model of communication. This might be done by various types of vocabularies, natural language, vocabulary derived from various parts of the document or by a more strictly defined vocabulary whose choice of terms is controlled.

For our purposes we will be concentrating on the controlled vocabulary used in the indexing and subsequent retrieval of information.

The definition of indexing given in the CARISPLAN Manual of Indexing Procedures is:

> "....the process of detailed subject analysis of a piece of literature, identification of the concepts contained, and the translation of these concepts into a pre-designated vocabulary."

In relation to the diagram on page 2, the quality of the indexing determines the amount of 'noise' in retrieval, the access which the user can gain to the information which is being given by the author, and the amount of information which is finally assimilated by the user.

1/ Shannon, Claude C. and Weaver, Warren. The mathematical theory of communication. Urbana: University of Illinois Press, 1949.

In assessment of the effectiveness of an information system, certain features which are common to all systems can be examined:

1. Recall which indicates the number of entries which are supplied by the system in reply to a search query using a particular indexing term.

2. Precision which is reflected by the number of entries which are directly relevant to the users' needs which can be located in reply to a search query using a particular indexing term.

The objective of indexing is to provide a satisfactory answer to the users' requests. The indexer undertakes a process of:

FAMILIARIZATION
ANALYSIS
CONVERSION OF TERMS TO INDEX TERMS

based on the anticipated needs of the system's main audience. The diagram on p. 14 of the CARISPLAN Manual of indexing Procedures identifies the stages of indexing showing that both indexer and searcher rely on use of the same vocabulary, which might or might not be controlled, for the translation of conceptual analysis into indexing terms.

The subject analysis undertaken for the indexing process is similar to that for abstracting but a further structuring is required to provide access points.

The main thrust is therefore toward determination of subject names.

The structure and usage of language does, however, present some potential problems in terms of making appropriate translations of concepts into index terms.

Synonyms require that there be determination of equivalence or levels of equivalence of terms which might in normal language be used interchangeably.

WAGES

SALARIES

INCENTIVES

are three terms which would need to have their equivalence determined, as would be the case with:

AUTOMATED

and

COMPUTERIZED

Common and scientific names can also be considered equivalent. These can in some cases be considered equivalent: in a general system, but in a very technical system, all the ingredients of SALT as used in everyday life might not be always exactly equivalent to:

SODIUM CHLORIDE.

Some synonyms also evolve from changes in use over time of terminology, so that

WIRELESS

might have been eliminated from a controlled vocabulary and replaced by

RADIO.

The cases of exact synonyms are, however, rare and on occasion the selection of one "synonym" over another can result in sacrifice of precision at the time of indexing, and high recall at the time of retrieval.

Homographs also have implications for the selection of indexing terminology, as the distinction between the meanings of such terms can usually be decided only by recognition of the context in which they are used.

LIME (fruit)

LIME (stone)

are examples of two terms which might need to be described in an S & T indexing system, and which would therefore require some description, or explanation to ensure that there is no ambiguity at the time of searching.

Number. A determination of whether singular or plural terms should be employed in the indexing vocabulary. For some terms, the difference between singular and plural is resolved simply by the addition of an 's' but in some cases the singular might be a completely different word.

CHILD

CHILDREN

In technical terminology the question of use of Compound terms is likely to be raised frequently during indexing and retrieval. Direct use of terms one would assume would lead to more spontaneous access:

ANIMAL NUTRITION

HUMAN NUTRITION

might be more quickly retrieved than:

NUTRITION, HUMAN

NUTRITION, ANIMAL

and, SOLAR HEATING

or, SOLAR POWER

might also be more easily located than:

HEATING, SOLAR

and, POWER, SOLAR

the differences are immediately evident. In the case of HUMAN NUTRITION, the user might require only information related to HUMANS, and might therefore be concerned at that point about other aspects of NUTRITION. In this case, searching directly for HUMAN NUTRITION would provide quicker access than searching by the inverted term NUTRITION, HUMAN.

Similarly, SOLAR HEATING and SOLAR POWER could be inverted without loss of meaning. One could anticipate, however, that there would be greater emphasis on the SOLAR aspect than was the case with NUTRITION, and that users who are searching for information on SOLAR HEATING might also be as interested in SOLAR POWER.

These two examples illustrate the use of compound terms, where the two subjects are equally valid as access points. It would be possible then to incorporate the discrete elements as entries in themselves and therefore to include all six terms:

NUTRITION

HUMAN

ANIMAL

SOLAR

POWER

HEATING

as entries in the indexing vocabulary.

There are other compound terms which consist of composite subjects, but cannot be as conveniently separated into indexing terms likely to lead to meaningful retrieval. Searches under:

TRANSFER

CYCLE

APPLIED

are likely to lead to false drops, while a search by:

TECHNOLOGY

is likely to lead to an excessively high recall.

SEMANTIC RELATIONSHIPS

The question of semantic relationships is also one likely to pose problems in the selection of indexing terms. It might be necessary to determine which levels of the generic relation should be treated.

For example, EXPANSION is a Narrower Term of DIMENSIONAL CHANGE, and a decision to use one or both of these terms would lead to different results in retrieval.

BASIS FOR SELECTION OF TERMS

While the choice of indexing terms presents questions in relation to synonyms, homographs, composite terms and semantic relationships, the general questions of exhaustivity and specificity need to be answered in relation to the treatment of documents.

EXHAUSIVITY

The recognition and translation in the indexing process of all meaningful concepts appearing in each document. By comparison with abstracting, the indexing terms chosen would be always retrievable and consequently exhausivity should ensure that all the information held in the system is retrievable.

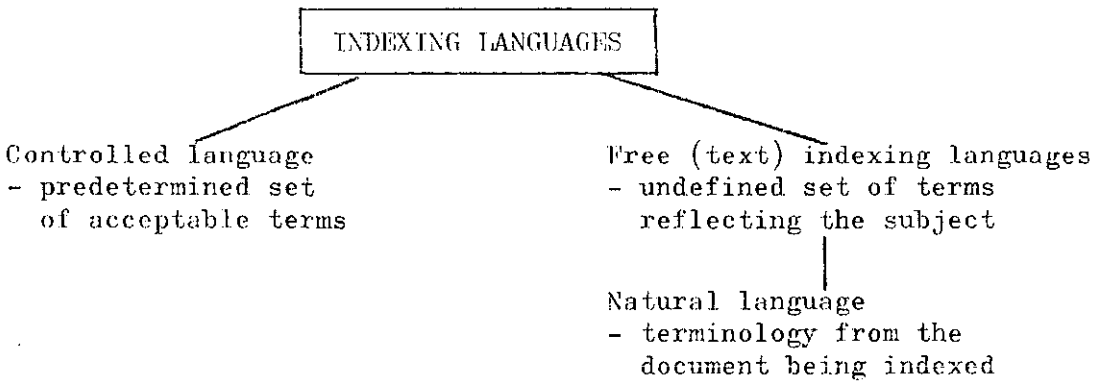
SPECIFICITY

The use of the most specific terms possible is again related to the choice of the narrowest term within the hierarchical relationship.

The use of both exhausivity and specificity naturally have implications for the level of precision and recall. Exhausivity would lead to higher recall and if linked to specificity, should contribute to more meaningful retreival.

TYPES OF INDEXING LANGUAGES

The translation of conceptual analyses into indexing terminology can be accomplished by various languages:



The basic questions raised earlier, apply more to some languages than to others and are also treated differently by each type of language.

CONTROLLED LANGUAGE generally gives much more specific directions to the indexer and the searcher, than is the case with free or natural language, and therefore, is likely to ensure much greater consistency in indexing, particularly in a networking environment.

Controlled language ranges from the pre-co-ordinate controls of languages such as the Library of Congress Subject Headings, to the basic simple term descriptors originally envisaged in the UNITERM System, with the majority of thesauri falling somewhere in between.

FREE LANGUAGE, as it has an infinite choice of words or is united only by language of the text, requires that the indexer answer the questions as they arise, and consistency could therefore only be achieved by the use of the more general guidelines rather than specific instructions.

Most computerized bibliographic systems and networks employ the post-co-ordinated indexing languages, but the trend in the newer systems is toward free text or natural language indexing. This is mainly because the free language which provides greater spontaneity in indexing can be searched very quickly by automated methods.

POST CO-ORDINATE INDEXING SYSTEMS

The basic design of post-co-ordinate systems envisaged the use of single, one concept terms with co-ordination of as many terms as might be required at the time of retrieval. The basic principles of the uniterm system are still applied in computerized bibliographic system.

1. Each document is first assigned a unique identifying number.

2. The document is then analysed and its subject represented by a number of indexing terms.

In both manual and automated systems, a file of subjects is held, and comparison of appropriate subjects is done at the time of retrieval.

In the basic un-co-ordinated file in a post-co-ordinate system, many entries make scanning tedious. But, in the case of a large number of entries, there can be a small number of headings.

Post-co-ordinate systems require files for:

- search keys
- document identification number

and these it seems can best be maintained simultaneously.

ITEM RECORDS

Records serially ordered by number

Storage

Catalogue cards by number

Punched cards

Magnetic tape

Disks/Diskettes

Characteristics

Less storage required

Detailed records

Infinite capacity but increased size reduces speed

Entire file needs to be searched for each query

Most suitable for applications such as SDI

TERM RECORDS

Records ordered by index in alphabetical order

Storage

Catalogue cards by index terms

Punched cards

Dual directory (computer produced)

Magnetic tape

Disks/diskettes

Characteristics

Small basic file

Detailed records

Infinite capacity but increased size reduces speed

Alphabetical access, quick result from query

More suitable for retrospective searching

THESAURI

The use of indexing languages in networking involves controlled, free text and natural language terminology, and although the use of free text and natural language is expanding, the indexing consistency aimed at in a decentralized network seems to be only readily achieved by the use of thesauri.

The distinction between thesauri and subject headings is difficult to demarcate completely. They are both types of controlled vocabularies which exist primarily:

- to provide standardization and control over synonyms quasi-synonyms and homographs
- to link semantically related terms
- to reduce or eliminate ambiguity of meaning
- to provide sufficient hierarchical structure to allow the conduct of generic searches

There are, however, certain characteristics which identify thesauri and distinguish them from "subject headings" and other classification schemes.

As thesauri are today more used in on-line searching there is a greater tendency toward more specific single concept descriptors, although pre-co-ordination is never completely eliminated.

Inverted terms are avoided by most thesauri, which either divide the composite term into its component parts, or enter the phrase in its conventional form.

Form and other sub-divisions are usually entered as terms in themselves separated from the main descriptor and therefore entries such as HYDRO-POWER-Bibliographies are rarely seen in thesauri.

Relationships between terms are broken down into:

Related term

Narrower term

Broader term

as compared to the "see also" directions of subject headings and similarly, the use of synonyms, quasi-synonyms and homographs is usually determined by USE and USE for instructions.

Relationships are in many thesauri also reflected by additional sequences supplementing the alphabetical which tends to be the first and most consulted.

In some cases as in the 1st edition of the SPINES Thesaurus, the structure of the relationships between terms are displayed by graphic displays.

One of the early activities in the design of a network is the examination of all relevant thesauri with a view to incorporating the most suitable one for use in the indexing process. The survey is usually based initially on the study of lists from thesaurus clearing houses such as Aslib or the University of Toronto School of Library and Information Science. Evaluation of those likely to be relevant would then be based on the requirements of users in relation to:

- the subject and the coverage by the thesaurus
- the type of literature covered by the network
- the quantity of material expected
- the type of information storage system - manual or automated
- the resources available mainly in terms of personnel to compile and update the thesaurus
- compatibility with other thesauri in related systems

The results of this evaluation would determine the need for construction of a new thesaurus or incorporation of an existing one into the network's indexing process.