DOCUMENT RESUME

ED 042 464 LI 002 083

TITLE	Guidelines for the Establishment and Development of Monolingual Scientific and Technical Thesauri for Information Retrieval.
INSTITUTION	United Nations Educational, Scientific, and Cultural Organization, Paris (France).
REPORT NO	SC-MD-20
PUB DATE	6 Jul 70
NOTE	14p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$0.80
DESCRIPTORS	Guidelines, Information Retrieval, *Lexicography, Lexicology, Subject Index Terms, *Thesauri

ABSTRACT

These guidelines for the establishment and development of monolingual scientific and technical thesauri for information retrieval are published in an attempt to lay the basis for compatability, both at the present and in the future, of thesauri that are being elaborated simultaneously in most of the disciplines of science, basic as well as applied. They are therefore, directed to all those who in the course of their careers come into contact with thesauri, either as users or as thesaurus compilers. Fourteen guidelines are present: the first four are of a general nature, the following seven deal with the establishment of thesauri, and the final three relate to the development of thesauri. These guidelines are specifically drafted in the English language and when applied to monolingual thesauri in other languages should be modified to take into consideration the attributes and uses of that particular language. (Author/MF)

Distribution: general

SC/MD/20
PARIS, 6 July 1970
Original: English

UNITED NATIONS EDUCATIONAL,
SCIENTIFIC AND CULTURAL ORGANIZATION

Guidelines for the Establishment and Development of Monolingual Scientific and Technical Thesauri for Information Retrieval

These guidelines were specifically drafted for the English language and when applied to monolingual thesauri in other languages they should be modified to take into consideration the attributes and uses of that particular language.

EXPLANATORY STATEMENT

At a time when the establishment of a World Science Information System is being seriously proposed* it is advisable to remember that the viability of any world system depends first and foremost on compatibility between its component parts.

These guidelines for the establishment and development of monolingual scientific and technical thesauri for information retrieval are published in an attempt to lay the basis for compatibility, both at present and in the future, of thesauri that are being elaborated simultaneously in most of the disciplines of science, basic as well as applied.

They are, therefore, directed to all those who in the course of their career come into contact with thesauri, either as users or as thesaurus compilers.

The first draft of the guidelines was prepared by the Unesco Secretariat. The third draft and this version were subsequently reviewed and studied by eminent and competent individuals and organizations and the relevant additions or corrections were made. Thus, these guidelines were presented to, and discussed by, the International Conference on the General Principles of Thesaurus Building in Warsaw, March 1970. The proposed changes are included in this version. May our collaborators accept anonymity along with our gratitude.

Fourteen guidelines are presented: the first four are of a general nature, the following seven deal with the establishment of thesauri, and the final three relate to the development of thesauri. Examples, where appropriate, are given on the right-hand margin of the text. By the word "thesaurus", as used in the present text, is meant 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 (auxiliary terms) as well as indicators of their relationships. Unlike classification schemes, the vocabulary does not necessarily use notations and categories.

A descriptor is an authorized and formulized term or symbol in a thesaurus, used to represent unambiguously the concepts of documents and queries.

Thesauri should be based on concepts and relationships which are internationally acceptable. Original and translated thesauri already exist in most of the major vehicular languages used in science and technology today. It is rare that any particular word can be translated univocally into another language without losing some shade of meaning in the process, but it is hoped that the application of these guidelines to monolingual thesauri will diminish the enormous difficulties encountered in the establishment of thesauri in different languages. These guidelines were originally drafted in English and when applied to monolingual thesauri in other languages, they should be modified to take into consideration the attributes and uses of that particular language (e. g. number of descriptors, VIII b).

* Joint Unesco-ICSU study on the feasibility of a World Science Information System (UNISIST) Final Report. Unesco, Paris, 1970.

'IN PRINT'

Thesauri can be used in many ways, and the structure of a thesaurus is intimately related to its proposed utilization. A thesaurus can be used merely as a word association list for helping indexers, or it can be considered as a transformation of the natural language into the information language.

Modern techniques in information science are nearly all based on the use of electronic computers and it is in this connexion that the use of thesauri is rapidly proliferating. It is this rapid proliferation which has brought the need for international guidelines to light and it was for this reason, too, that Unesco recently encouraged (helping in the establishment of one) the work of two clearing-houses dealing with thesauri. These clearing-houses are located at the Bibliographic Systems Center, School of Library Science, Case Western Reserve University, Cleveland, Ohio 44106, United States of America, and at the Centralny Instytut Informacji Naukowo-Technicznej i Ekonomicznej, Al. Niepodległości 188, Warsaw, Poland for English and languages other than English respectively.

GENERAL

I. ADVISABILITY OF A PILOT RUN

Before establishing a thesaurus on a definitive basis it is strongly recommended that a practical test, based on a restricted number of documents dealing with a small area of the domains to be ultimately covered, be carried out. This pilot run, based on tentatively structured terms, should show up the more adequate methods of descriptor selection and thesaurus display applicable to the case under consideration. The results of this test should be critically commented upon by as many people as feasible, including information scientists and indexers as well as subject specialists and users.

II. NECESSITY OF A DESCRIPTIVE INTRODUCTION TO THE THESAURUS

No thesaurus should be presented without a comprehensive introduction which states clearly the purpose and structure of the thesaurus, and the domains covered by it. The rules followed in its establishments should be presented in a condensed form. This is particularly true of the methods and sources used in the selection, form and avoidance of ambiguity of the descriptors (see VI, VII, VIII). The method of presenting the thesaurus as well as the rules for alphabetization and punctuation, whenever applicable, should be explicitly stated.

Most important of all, the rules for using the thesaurus and its limits of applicability should be

The total number of descriptors, non-descriptors, identifiers, (see VI), hierarchical chains (see X (b)) and related concepts (see X (c) should be itemized.

III. NECESSITY OF INDEXES

Every thesaurus, regardless of its mode of presentation (see XI) should contain an alphabetical union list of each individual unstructured term whether issued separately as a supplement or together with the main thesaurus as an annex. Permutation indexes may also be used.

It may be useful in the case of multidisciplinary thesauri to present, in addition, indexes in which the descriptors are grouped by discipline.

IV. NOTIFICATION OF INTENT

The appropriate clearing-house (see above) should be notified of the intention to construct a thesaurus, as well as when the thesaurus is first published or disseminated. This information should be channelled through the national organization dealing with thesauri, where and when such an entity exists.

The same applies for further editions. If at all possible, a copy of the thesaurus, complete with the introduction and indexes should be sent to the clearing-house in question. The fact of notification should be mentioned in the introduction.

ESTABLISHMENT

V. CHECK WITH CLEARINGHOUSE TO AVOID DUPLICATION

Before commencing work on the establishment of the thesaurus, it is advisable to ascertain whether others covering that particular domain or a neighbouring one are available.

This is best done by addressing a query to the two clearinghouses mentioned above. It may be found advisable to go ahead with the compilation of a particular thesaurus in spite of the existence of a similar one. In this case the reasons for proceeding and the differences with the earlier thesaurus should be clearly stated in the introduction.

VI. SELECTION OF DESCRIPTORS

The selection of descriptors should begin only after the general structure of the thesaurus has been agreed upon. It should be carried out, preferably, by people who have both a good knowledge of the subject to be treated, and previous experience in indexing or classification. The use of internationally recruited teams for the [illegible] mechanical processing, introduction of a new domain e. g. interdisciplinary areas for which no previous classification schemes existed, existence of well-defined group of users and subject specialists, extensive literature).

Acoustical Holography
Brain Research

Descriptors, in general, consist of terms related to discrete concepts encountered in the subject field under consideration and in pertinent marginal areas. A more specific class of thesaurus terms known as "identifiers" may sometimes be used.

Descriptors should succinctly summarize concepts in as few words as possible, preferably one. Grammatical connexions such as prepositions or articles should be avoided whenever possible.

Identifiers constitute a special type of thesaurus terms which are not reciprocally cross-referenced (see XI) and which serve the purpose of providing additional indexing depth. For instance, identifiers might include individual trade names, geographical locations, equipment, nomenclature, code names etc.

IRELAND
NT DUBLIN
DUBLIN

Since they are not reciprocally cross-referenced, identifiers need not necessarily appear in the thesaurus display, but may be listed separately, in addition to appearing in the Union List (see III above).

Four distinct steps intervene in the selection of descriptors: collection, verification, evaluation, and choice.

(a) Collection

It is almost impossible to make a comprehensive collection of candidate descriptors by thinking of an alphabetical list. By envisaging descriptors in groups, thought associations between them give rise to many candidates. Potential users and subject specialists as well as internationally or nationally standardized technical dictionaries should be consulted; terms should be chosen from the current literature; existing word lists or classification schemes should be culled and may be expanded or compressed appropriately. Scientific and technical dictionaries and glossaries, both multilingual and monolingual constitute a prolific source of descriptors (see page 14).

(b) Verification

With all methods of assembly, the authenticity of the selected descriptors should be verified by consulting dictionaries, other indexing or standardized vocabularies, current usage in the literature and especially the opinion of subject specialists. Obsolete terminology should not be included, or if so only as forbidden terms (see X (a) below).

One of the more appealing attributes of a thesaurus is its ability to assimilate immediately the neologisms and special [illegible] that proliferate in expanding fields of basic and applied re-

(c) Evaluation

In evaluating the utility of candidate descriptors, reference should be made to their: (1) frequency as encountered in the literature or in the existing stocks of information; (2) anticipated incidence in retrieval inquiries; (3) relationship to descriptors already accepted; (4) appropriateness and authenticity as current terminology in the discipline concerned; (5) effectiveness and expediency in connoting and denoting the particular concept. None of these factors should be considered independently and particular attention should be paid to areas of peripheral interest where the exhaustivity and specificity required of the descriptors are not the same as for the core subject.

(d) Choice

In all cases, descriptors should be selected for inclusion in the thesaurus on the basis of their estimated effectiveness for retrieval purposes and their measureable significance in the material to be indexed.

VII. METHODS OF AVOIDING AMBIGUITY

In compiling a thesaurus, difficulties are encountered with descriptors which have more than one accepted meaning or whose meaning in a given context is different to that commonly encountered. In such cases the required meaning may be brought out by the use of the following methods:

(a) Compound expressions

Although descriptors are preferably self-contained, single term concepts, the use of modifying expressions to make clear the different meanings associated with a given term is necessary in certain cases. For the method of entering the resulting compound expression, (see IX (a) below).

LATENT HEAT

(b) Qualifiers for homonyms

The various forms of homonyms may be distinguished by the use of qualifying expressions placed between parentheses immediately after the homonym. Other homonyms should not be used as parenthetic qualifiers.

BEAMS (ELECTRO-MAGNETIC)
BEAMS (STRUCTURAL)

(c) Scope notes

A scope note is a brief explanation which may accompany the descriptor in the thesaurus display, but does not form part of the descriptor. It indicates the way in which the descriptor should be used; it need not necessarily consist of a dictionary definition. Scope notes are sometimes used to restrict the usage of a descriptor. They should always be used in connexion with abbrevia-

*DOCUMENTATION
The process of storing and retrieving information in all fields of learning.
*DOCUMENTATION
The volume of documents

VIII. FORM OF DESCRIPTORS

(a) Word form

Once it has been decided to include a given term in the thesaurus, care should be taken to ensure that the word form used adequately conveys the exact meaning intended.

(i) Spelling: the most widely accepted spelling of the word should be used. Cases arise, particularly in English, due to varying usage on different sides of the Atlantic, where more than

* For instance, in three different thesauri. If these three meanings were in the same thesaurus, they would require qualifiers in order to make them unique.

one spelling of a word is accepted, in which case both forms of the word should be included in the thesaurus. In these cases the preferential cross-reference should be employed (see X (a) below). Alternatively, a well-established dictionary can be chosen to act as arbitrator whenever this problem arises.

SULFUR
SULPHUR

(ii) Translation: many current technical terms have arisen by translation from other languages, but sometimes a modern foreign language or Latin term is incorporated into the specialized vocabulary for a particular subject. When both the foreign language term and its putative translation coexist, they should both be included in the thesaurus and cross-referenced preferentially.

BRAKING RADIATION
BREMSSTRAHLUNG

(iii) Transliteration: the problem is further complicated when the foreign language in question is written in a different alphabet. This is particularly true in the case of identifiers (see VI above). The transliteration standards* recommended by the International Organization for Standardization should be used whenever applicable. Wherever a choice exists, the transliteration which does not employ diacritical marks should be selected (see (e) below).

SATELLITE
SPUTNIK

(b) Noun form

The descriptor should be in the form of a noun or that part of the verb which is grammatically equivalent.

The gerund in English

(c) Number

In general, the plural form should be used for descriptors, particularly when generic terms are involved. The singular form is used for specific material or property terms, process terms, proper names and disciplinary areas. Sometimes the singular and plural forms of a word denote different concepts, in this case both should be entered.

FORCES
HEATING
PALYNOLOGY
TEAK

WOOD
WOODS

(d) Abbreviations and acronyms

Abbreviated word forms should be used only when their meaning is internationally established. Both abbreviated and unabbreviated forms should be displayed and cross-referenced preferentially. Sometimes the necessity for limiting the length of the descriptor

(e) Character set

Since the majority of scientific and technical thesauri now being established will probably be used in connexion with electronic computers, it is advisable to use only the upper case format for the descriptors. Diacritical marks should be avoided for the same reason.

The need for these restrictions will probably disappear in the near future as the fruits of technical advances become more widely distributed, and computer manufacturers pay more heed to the exhortations of information scientists to lower the costs of peripheral equipment.

* See page 14

As mentioned in (d) above, the eventual use of a computer may entail the limiting of the number of characters that a descriptor may have.

(f) Special characters and numerals

The only special characters allowed in descriptors are left and right parentheses and unavoidable hyphens. (Fullstops may sometimes be used (see IX (b) below)). Any other non-alphanumeric symbols should be confined to scope notes, always within the limits of machine character availability. If the descriptors contain numeric elements, arabic numerals should be used. The position of the numerals should follow normal usage. Rules must be established for the treatment of subscript and superscript numerals.

MERCURY (PLANET)

In the particular case of data retrieval thesauri, the stroke ("/") may sometimes be found necessary.

EH/M

IX. METHODS OF ENTERING DESCRIPTORS IN THE THESAURUS

(a) Syntax

Compound expressions consisting of two or more words should be listed preferably by direct entry i.e. not artificially inverted. This is especially true for descriptors: for forbidden terms, this recommendation may be relaxed. Evidently this does not apply when a permuted or key word in context type of multiple entry is used. Inverted entries may be used provided they are preferentially cross-referenced (see X (a)). When a qualifier between brackets (see VII (b) above) forms part of the descriptor it is advisable to enter the qualifier on its own with a preferential cross-reference to the complete descriptor.

ELECTRICAL POWER
not
POWER, ELECTRICAL

(b) Punctuation

Punctuation marks should not appear in the descriptors. As stated [illegible] terms. Where punctuation marks are omitted, it is advisable to include them in full in the scope notes.

(c) (i) Specialized vocabularies

Certain fields have highly specific systems of nomenclature, or well-established standardized technical vocabularies. Whenever an internationally agreed nomenclature exists, it should be used.

(ii) Specific names

The proliferation of unrelated specific names would tend to convert the thesaurus into a simple list of identifiers which would be self-defeating. It is therefore recommended that the names of unrelated specific entities be avoided as much as possible.

(iii) Specific items

Descriptors representing generic, functional or structural concepts can be co-ordinated to denote specific items, while be retaining the

property of being cross-referenced, they fulfil the structural needs of thesaurus elements.

(d) Alphabetization

Where appropriate, one of the following alphabetization methods may be followed:

(i) letter by letter
(ii) word by word
(iii) computer sort

The selection of the method of alphabetization depends on all the factors affecting the thesaurus under construction. i.e. the size and structure of the domains covered by the thesaurus, the availability of machine processing, the kind of hardware available, etc. In all cases the alphabetization rules should be clearly and explicitly drawn up before any kind of ordering is attempted.

(e) Synonyms and quasi-synonyms

It is rare that two or more candidate descriptors can be considered as true synonyms. When one candidate descriptor must be searched every time that another is searched, they may be treated as synonyms. Descriptors that overlap significantly or represent different aspects of the same property may be considered quasi-synonyms. Antonyms should be similarly treated. When all the synonyms, quasi-synonyms or antonyms are included in the thesaurus display the preferential cross-reference should be used (see X (a) below).

COLUMBIUM/NIOBIUM

HEREDITY/GENETICS

HARDNESS/SOFTNESS

X. INTERRELATIONSHIPS BETWEEN DESCRIPTORS

The most important function of a thesaurus is to serve as a tool for information retrieval. Therefore it should bring into evidence the interrelationship between individual descriptors. These can be expressed by several means. If codes are used to indicate these relationships, their meaning should always be made clear.

This is necessary for the homogeneity of the thesaurus and for "book-keeping" purposes.

(a) Preferential

This reference is employed to refer from a forbidden term to a descriptor and vice versa. It is used when the meaning of descriptors overlaps substantially; where different spellings of the same word exist; for synonyms, quasi-synonyms and antonyms and, in general, wherever a choice has been made between a number of descriptors, all of which are included in the thesaurus display.

Common codes in English are:
use/includes
use (USE)/used for (UF)

ALCOHOLS
USE ALKANOLS
ALKANOLS
UF ALCOHOLS

(b) Hierarchical

Hierarchical relationships are used to exhibit relative degrees of specificity within a category of descriptors all of which belong to

Common codes in English are:
broader term (BT)/
narrower term (NT)

a particular generic group. This relationship is not based upon the possible use or application of an entity, but on the position of the descriptor within a given class of concepts. Note that certain terms may be members of more than one hierarchical chain. Where any hierarchy has more than two levels the cross-references for all levels should be completed for each descriptor. The kinds of hierarchical relationships which it is desirable to indicate depend on the structure of the subject field of the thesaurus. In general, all concepts which are sub-divisions of a broader concept should form part of a hierarchical chain.

specific to/generic to

CALCULUS
NT INTEGRAL CALCULUS

INTEGRAL CALCULUS
BT CALCULUS

Genus-species in zoology

Whole-part

Subordinate concepts

(c) Affinitive

The affinitive relationship is employed to refer from a descriptor to others that are closely related in concept but are neither consistently hierarchically nor preferentially related. This relationship may be based on usage, application, physical proximity, etc.

Common codes in English are:
related term (RT)

also see
EDUCATION
RT LEARNING

LEARNING
RT EDUCATION

XI. PRESENTATION OF THESAURUS

It is recommended that a thesaurus be presented in one or more systematical displays and alphabetical listings.

(a) Systematical listing

Systematical listing refers to that form of thesaurus display in which descriptors are first of all grouped in general class categories within each of which the interrelationships between the descriptors, particularly the hierarchical relationships, are as self-contained as possible. Full use should be made of recorded experience in the field of classification when establishing the membership of the various facets.

Some descriptors may appear in more than one category but this should occur only when either the descriptor is accompanied by a parenthetical qualifier or when cross-references are used.

[redacted]

[redacted] of display with (c) below gives rise to a kind of structured alphabetical list which probably combines to the fullest extent the advantages of both.

(b) Graphic display

Perhaps the most subtle mode of presentation of thesauri is to display the descriptors and the relationships between them graphically. Although this can be done multi-dimensionally, for instance by taking two dimensions for each facet of a multi-faceted thesaurus the more current methods are two-dimensional.

One such system consists of arranging the descriptors in semantic groups, assigning a gridded sheet to each group and giving fixed positions to each descriptor with respect to the horizontal and vertical axes, thus defining co-ordinates.

Interrelationships between descriptors are then shown by means of arrows. Associative relationships are denoted by bi-directional arrows. Hierarchical relationships are shown by uni-directional arrows always pointed to the more specific descriptor. Preferential relationships may be indicated by brackets with the arrows leaving or arriving at the preferred term.

It is understood that a descriptor may belong to several groups. The optimal size of each group appears to lie between 30 and 40. As before, an alphabetical listing should be given in annex showing the semantic group(s) to which each descriptor belongs. Which mode of presentation is selected will depend on the use to which the particular thesaurus will be put.

The latter two types of display lend themselves more easily to translation. A rather particular type of thesaurus is the following.

(c) Alphabetical listing

The great advantage of an alphabetical listing is that the introduction and correct positioning of new descriptors is very easy. On the other hand, it is extremely difficult to introduce structure into a strictly alphabetical list. For instance, synonyms come more readily to mind if we think of a particular category as a whole rather than individual descriptors plucked at random from an alphabetical list. It should be remembered that a particular alphabetical order is only applicable in one language. Permuted alphabetical lists may also be used.

DEVELOPMENT

XII. PERIODIC VERIFICATION OF USEFULNESS OF INDIVIDUAL DESCRIPTORS

At least for the first few years, if not permanently, after the establishment of a thesaurus, a check should be kept on the frequency with which particular descriptors are utilized, both for indexing and retrieval purposes. Periodic verification should ensure that certain descriptors neither interfere with, nor duplicate one another. On all occasions in which a search does not locate the desired information or the amount of information suspected of being in the collection, a critical appraisal of the descriptors which were, or should have been used, ought to be carried out.

XIII. ELIMINATION OF DESCRIPTORS

If it is found that any descriptor is being used very infrequently, care should be taken to ensure that the infrequency of usage is not due purely to the lack of documents related to that particular concept. It may either be eliminated from the thesaurus or replaced by another more common term. Complete elimination should occur ideally only when that particular descriptor has never been used, [illegible]

[illegible] preferential relationship is not used, the date of introduction of a new descriptor into the thesaurus should be noted since, prior to that date, indexers were not authorized to use that term.

The procedure to be followed when a particular descriptor is over or under used depends to a certain extent on the search strategy employed in retrieval. If the least specific descriptor is searched for last, it may not be worth while to eliminate it.

XIV. CHOICE OF NEW DESCRIPTORS

Indexers and users should constantly be on the look-out for new candidate descriptors which may represent either new concepts or different facets of old concepts. If possible, the descriptor should be used on a trial basis by indexers for some time before becoming a definite addition to the thesaurus.

The frequency of occurrence of such candidate descriptors both as indexing and retrieval terms is a good indication of their future usefulness. If it is decided to add a new descriptor, the interrelationships with all the pre-existing descriptors should be identified and introduced in the appropriate places.

Definite additions should not be introduced singly as this causes confusion among the users of the thesaurus. New descriptors should be saved up and introduced by batches, either as "additions to the thesaurus" or on the occasion of a new edition of the thesaurus. This does not preclude their use by indexers. There should exist a central authority which examines all the suggestions received and issues a final verdict on the acceptability or otherwise of the possible new additions.

It should always be remembered that a thesaurus is never completed, its size and shape being a function of time.

Any comments on the above text may be sent to the Division of Scientific Documentation and Information, Unesco, Place de Fontenoy, 75 - Paris VII, France.

LIST OF ISO RECOMMENDATIONS RELATED TO THESE GUIDELINES

ISO/R 9	"International system for the transliteration of slavic Cyrillic characters" 2nd edition.
ISO/R 233	"International system for the transliteration of Arabic characters".
ISO/R 259	"Transliteration of Hebrew".
ISO/R 704	"Naming principles".
ISO/R 843	"International system for the transliteration of Greek characters into Latin characters".
ISO/R 860	"International unification of concepts and terms".
ISO/R 919	"Guide for the preparation of classified vocabularies".
ISO/R 1087	"Vocabulary of terminology".
ISO/R 1149	"Layout of multilingual classified vocabularies".
ISO/DR 1951	"Lexicographical symbols, particularly for use in classified defining

Sources for dictionaries and glossaries:

Bibliography of interlingual scientific and technical dictionaries, 5 ed. Paris. Unesco, 1969. 25[illegible] p.

Bibliography of monolingual scientific and technical glossaries, Vol. I: National Standards 1955, 219 p. Vol. II: Miscellaneous Sources, 1959, 146 p. Paris, Unesco.

(Supplements published in Babel, International Journal of Translation published by the International Federation of Translators with the assistance of Unesco. Avignon, France.)

Bibliographic Bulletin of the Clearinghouse at CINTE, 1969, Warsaw, CINTE, 1969. 140 p. (Annual supplements are planned.)

Bibliographic Systems Center Subject Index, Case Western Reserve University, Cleveland, Ohio, U.S.A., 1969. (Computer print-out.)

Some national standards institutions publish extensive unilingual and sometimes bilingual technical vocabularies.