

BS ISO 25964-1:2011



BSI Standards Publication

Information and documentation — Thesauri and interoperability with other vocabularies

Part 1: Thesauri for information retrieval



...making excellence a habit.
AVA Documents™
Document Sharing Hub

National foreword

This British Standard is the UK implementation of ISO 25964-1:2011. It supersedes BS 8723-2:2005 and DD 8723-5:2008, which are withdrawn. Together with BS ISO 25964-2, it supersedes BS 8723-1:2005 and BS 8723-4:2007, which will be withdrawn on publication of BS ISO 25964-2.

The UK participation in its preparation was entrusted to Technical Committee IDT/2/2, Information description, source identification and ,indexing.

A list of organizations represented on this committee can be obtained on request to its secretary.

This publication does not purport to include all the necessary provisions of a contract. Users are responsible for its correct application.

© BSI 2011

ISBN 978 0 580 58905 8

ICS 01.140.20

Compliance with a British Standard cannot confer immunity from legal obligations.

This British Standard was published under the authority of the Standards Policy and Strategy Committee on 30 September 2011.

Amendments issued since publication

Date	Text affected
------	---------------

INTERNATIONAL STANDARD

BS ISO 25964-1:2011

ISO
25964-1

First edition
2011-08-15

Information and documentation — Thesauri and interoperability with other vocabularies —

Part 1: Thesauri for information retrieval

*Information et documentation — Thésaurus et interopérabilité avec
d'autres vocabulaires —*

Partie 1: Thésaurus pour la recherche documentaire



Reference number
ISO 25964-1:2011(E)

AVA **Documents**
Document Sharing Hub
© ISO 2011



COPYRIGHT PROTECTED DOCUMENT

© ISO 2011

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword	v
Introduction.....	vi
1 Scope	1
2 Terms and definitions	1
3 Symbols, abbreviated terms and other conventions.....	12
4 Thesaurus overview and objectives.....	15
4.1 Overall objective.....	15
4.2 Vocabulary control and its purpose	16
4.3 Paradigmatic versus syntagmatic relationships.....	16
4.4 Types of paradigmatic relationship.....	17
5 Concepts and their scope in a thesaurus	18
5.1 Conceptual basis	18
5.2 Scope notes	20
5.3 Reciprocal scope notes	21
6 Thesaurus terms.....	21
6.1 Form of terms	21
6.2 Clarification and disambiguation of thesaurus terms	21
6.3 Grammatical form of terms.....	23
6.4 Capitalization, punctuation and special characters	26
6.5 Singular or plural forms.....	27
6.6 Selection of the preferred form.....	30
7 Complex concepts.....	37
7.1 General	37
7.2 The nature of compound terms	38
7.3 Deciding whether or not to admit a complex concept.....	39
7.4 How to split a complex concept.....	43
7.5 Retention of constituent concepts	43
7.6 Consistency in the treatment of complex concepts	44
7.7 Order of words in multi-word terms	44
8 The equivalence relationship, in a monolingual context	44
8.1 General	44
8.2 Synonyms.....	45
8.3 Quasi-synonyms.....	48
8.4 Specific terms subsumed in a broader concept	48
8.5 Representation of complex concepts by a combination of terms	49
9 Equivalence across languages	50
9.1 General	50
9.2 Degrees of equivalence	51
9.3 Typical problems and solutions	52
9.4 Representation of cross-language equivalence between preferred terms	57
9.5 Cross-language equivalence between non-preferred terms.....	57
10 Relationships between concepts.....	57
10.1 Introduction.....	57
10.2 The hierarchical relationship	58
10.3 The associative relationship	63
10.4 Customized relationships.....	67

11	Facet analysis	68
12	Presentation and layout	70
12.1	General.....	70
12.2	Alternative display styles.....	71
12.3	Presentation and layout of multilingual thesauri	80
12.4	Language and character encoding issues	85
13	Managing thesaurus construction and maintenance	88
13.1	Planning a thesaurus	88
13.2	Early stages of compilation	90
13.3	Construction.....	91
13.4	Introduction to the thesaurus.....	93
13.5	Dissemination	93
13.6	Updating	95
14	Guidelines for thesaurus management software	98
14.1	General.....	98
14.2	Size and character limitations	98
14.3	Relationships between terms and between concepts	99
14.4	Notes applying to terms or concepts	100
14.5	Codes and notation	100
14.6	Node labels.....	100
14.7	Status of languages.....	100
14.8	Data import/export.....	101
14.9	Editorial navigation and support.....	102
14.10	Editorial safeguards	102
14.11	Housekeeping tools.....	103
15	Data model.....	103
15.1	General.....	103
15.2	Notes on the model.....	105
15.3	Tabular presentation	109
16	Integration of thesauri with applications	115
16.1	Introduction	115
16.2	Interoperability needs for thesauri.....	116
16.3	Integration with indexing and searching applications.....	116
17	Exchange formats.....	118
18	Protocols	119
18.1	General.....	119
18.2	Purposes and use cases	119
18.3	Application environment and architecture	120
18.4	Thesaurus-specific protocols.....	120
18.5	General-purpose web database protocols used with thesauri	120
	Annex A (informative) Examples of displays found in published thesauri	122
	Annex B (informative) XML Schema for data exchange.....	139
	Bibliography.....	140
	Index.....	144
	Table 1 — Symbols and abbreviations	13
	Table 2 — English language tags and their equivalents in other languages	14
	Table A.1 — Tags used in Inspec Thesaurus alphabetical display	122
	Figure 1 — Paradigmatic and syntagmatic relationships.....	17

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 25964-1 was prepared by Technical Committee ISO/TC 46, *Information and documentation*, Subcommittee SC 9, *Identification and description*.

This first edition of ISO 25964-1 cancels and replaces ISO 2788:1986 and ISO 5964:1985, which have been merged and technically revised. Clauses 1 to 13 of this part of ISO 25964 correspond broadly to the content of ISO 2788:1986 and ISO 5964:1985. The remaining clauses cover new material.

ISO 25964 consists of the following parts, under the general title *Information and documentation — Thesauri and interoperability with other vocabularies*:

— *Part 1: Thesauri for information retrieval*

The following parts are under preparation:

— *Part 2: Interoperability with other vocabularies*

This part of ISO 25964 covers the development and maintenance of thesauri, both monolingual and multilingual, including formats and protocols for data exchange.

ISO 25964-2 will cover interoperability between different thesauri and with other types of structured vocabulary, such as classification schemes, name authority lists, ontologies, etc., not previously covered in any International Standard.

Introduction

Today's thesauri are mostly electronic tools, having moved on from the paper-based era when thesaurus standards were first developed. They are built and maintained with the support of software and need to integrate with other software, such as search engines and content management systems. (For example, data from the thesaurus database might need to be presented in combination with the number of postings found by a search application.) Whereas in the past thesauri were designed for information professionals trained in indexing and searching, today there is a demand for vocabularies that untrained users will find to be intuitive, and for vocabularies that enable inferencing by machines.

ISO 25964 makes the transition that is needed in order to be compatible with the world of electronic information management. However, this part of ISO 25964 retains the assumption that human intellect is usually involved in the selection of indexing terms and in the selection of search terms. If both the indexer and the searcher are guided to choose the same term for the same concept, then relevant documents will be retrieved. This is the main principle underlying thesaurus design, even though a thesaurus may also be applied in situations where computers make the choices.

Efficient exchange of data is a vital component of thesaurus management and exploitation. This part of ISO 25964 therefore includes recommendations for exchange formats and protocols. Adoption of these will facilitate interoperability between thesaurus management systems and other computer applications, such as indexing and retrieval systems, that will utilize the data.

This part of ISO 25964 covers development and maintenance of thesauri rather than how to use them in indexing. Where multilingual issues and examples are addressed, efforts have been made to cover as wide a selection of languages as possible, consistent with clarity and comprehensibility.

Thesauri are typically used in post-coordinate retrieval systems, but may also be applied to hierarchical directories, pre-coordinate indexes and classification systems. Increasingly, thesaurus applications need to mesh with others, such as automatic categorization schemes, free-text search systems, etc. ISO 25964-2 will address additional types of structured vocabulary (such as classification schemes, name authority lists, ontologies, etc.) and give recommendations to enable interoperation of the vocabularies at all stages of the information storage and retrieval process.

Information and documentation — Thesauri and interoperability with other vocabularies —

Part 1: Thesauri for information retrieval

1 Scope

This part of ISO 25964 gives recommendations for the development and maintenance of thesauri intended for information retrieval applications. It is applicable to vocabularies used for retrieving information from all types of information resources, irrespective of the media used (text, sound, still or moving image, physical object or multimedia) including knowledge bases and portals, bibliographic databases, text, museum or multimedia collections, and the items within them.

This part of ISO 25964 also provides a data model and recommended format for the import and export of thesaurus data.

This part of ISO 25964 is applicable to monolingual and multilingual thesauri.

This part of ISO 25964 is not applicable to the preparation of back-of-the-book indexes, although many of its recommendations could be useful for that purpose.

This part of ISO 25964 is not applicable to the databases or software used directly in search or indexing applications, but does anticipate the needs of such applications among its recommendations for thesaurus management.

2 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

2.1

array

group of **sibling concepts** (2.52)

EXAMPLE

In the following, the sibling concepts outerwear and underwear form an array within the concept “clothing”.

- clothing
 - outerwear
 - overcoats
 - underwear

2.2

associative relationship

relationship between a pair of **concepts** (2.11) that are not related hierarchically but share a strong semantic connection

2.3

broader term

preferred term (2.45) representing a **concept** (2.11) that is broader than the one in question

NOTE The scope of the narrower concept falls completely within the scope of the broader. The relationship between the two is commonly indicated with the tag BT. For more explanation see 10.2.1.

2.4

characteristic of division

attribute by which a **concept** (2.11) can be subdivided into an **array** (2.1) of narrower **concepts** (2.11), each having a distinct value of that attribute

cf. **facet analysis** (2.21), **node label** (2.38)

EXAMPLE

In the following, age group is the characteristic of division applied to the concept of people:

people
(people by age group)
children
youths
adults

2.5

classification

classifying

activity involving the components of grouping similar or related things together; separating dissimilar or unrelated things; and arranging the resulting groups in a logical and helpful sequence

2.6

classification scheme

schedule (2.49) of **concepts** (2.11) and pre-coordinated combinations of **concepts** (2.11), arranged by **classification** (2.5)

NOTE A classification scheme often also includes an index.

2.7

coined term

new **term** (2.61) created to express a **concept** (2.11) for which no suitable **term** (2.61) exists in the required language

NOTE For a further explanation and examples, see 6.6.5 and 9.3.3.3

2.8

compound equivalence

relationship or mapping in which one **term** (2.61) or **concept** (2.11) in one context is represented by two or more **terms** (2.61) or **concepts** (2.11) in another

2.9

compound term

term (2.61) that can be split morphologically into separate components

EXAMPLE

In English:

“copper mines” can be split into “copper” and “mines”; “lawnmowers” can be split into “lawns” and “mowers”

In French:

“mine de cuivre” can be split into “mine” and “cuivre”; “biodiversité” can be split into “biologie” and “diversité”

NOTE Compound terms can be multi-word terms, or can consist of only one word.

2.10

computer application

computer program or set of programs that provides high-level processing related to a specific user need

NOTE In ISO 25964, a computer application is sometimes referred to as an application.

2.11

concept

unit of thought

NOTE Concepts can often be expressed in a variety of different ways. They exist in the mind as abstract entities independent of terms used to express them. They range from the very simple, e.g. “child”, to the very complex, e.g. “child protection legislation”.

2.12

controlled vocabulary

prescribed list of **terms** (2.61), headings or codes, each representing a **concept** (2.11)

NOTE Controlled vocabularies are designed for applications in which it is useful to identify each concept with one consistent label, for example when classifying documents, indexing them and/or searching them. Thesauri, subject heading schemes and name authority lists are examples of controlled vocabularies.

2.13

cross-language equivalence

equivalence relationship (2.18) between **terms** (2.61) representing the same **concept** (2.11) in different languages

2.14

data model

abstract model that describes how data is represented and used

NOTE The data model in this part of ISO 25964 provides a generic definition of thesaurus structure and semantics. It can be used as the basis for defining a database model or an exchange format for thesauri.

2.15

document

any resource that can be classified or indexed in order that the data or information in it can be retrieved

NOTE This definition refers not only to written and printed materials in paper or microform versions (for example, conventional books, journals, diagrams, maps), but also to non-printed media such as machine-readable and digitized records, Internet and intranet resources, films, sound recordings, people and organizations as knowledge resources, buildings, sites, monuments, three-dimensional objects or realia; and to collections of such items or parts of such items.

2.16

entry term

lead-in term

term (2.61) provided in a **controlled vocabulary** (2.12), not for direct use in **metadata** (2.33), but for the purpose of guiding the user to another **term** (2.61) that can be used as a category label, subject heading or **preferred term** (2.45)

NOTE Entry terms occurring in a thesaurus are generally known as non-preferred terms.

2.17

equivalence mapping

mapping that states that the **concept** (2.11) in the target vocabulary is considered identical in scope to the **concept** (2.11) in the source vocabulary

cf. **equivalence relationship** (2.18)

2.18

equivalence relationship

relationship between two **terms** (2.61) in a **thesaurus** (2.62) that both represent the same **concept** (2.11)

NOTE In ordinary discourse, terms that are quasi-synonyms may represent slightly different concepts. After inclusion in the thesaurus, however, the equivalence relationship clarifies that both are regarded as representing the same concept. When two or more such terms are in the same language within a monolingual or multilingual thesaurus, one of them is designated a preferred term and the other(s) non-preferred term(s); when two or more such terms are in the different languages of a multilingual thesaurus, each of them may be a preferred term in its own language respectively, and the relationship is known as cross-language equivalence.

2.19

exchange format

machine-readable format for representing information that is intended to facilitate exchange of the information between different applications

NOTE The exchange format for a thesaurus often uses a markup language based on a standard such as XML (Extensible Markup Language) [63][64][65][66], and is based on a data model for thesauri. While the data model provides a generic description of thesaurus structure and semantics, the exchange format expresses this in a formal language for the purpose of exchanging thesauri.

2.20

facet

grouping of **concepts** (2.11) of the same inherent category

EXAMPLE 1 Animals, mice, daffodils and bacteria could all be members of a living organisms facet.

EXAMPLE 2 Digging, writing and cooking could all be members of an actions facet.

EXAMPLE 3 Paris, the United Kingdom and the Alps could all be members of a places facet.

NOTE Examples of high-level categories that can be used for grouping concepts into facets are: objects, materials, agents, actions, places and times.

cf. **node label** (2.38)

2.21

facet analysis

analysis of subject areas into constituent **concepts** (2.11) grouped into **facets** (2.20), and the subdivision of **concepts** (2.11) into narrower **concepts** (2.11) by specified characteristics of division

2.22

facet indicator

notational device that indicates the start of a new **facet** (2.20) within a synthesized compound **notation** (2.40)

NOTE Examples of facet indicators are the 0 in the Dewey Decimal Classification, and parentheses and quotation symbols in the Universal Decimal Classification. In the past, the term facet indicator has been used as synonymous with node label but that usage is deprecated by ISO 25964, to avoid confusion.

2.23

hierarchical relationship

relationship between a pair of **concepts** (2.11) of which one has a scope falling completely within the scope of the other

cf. **broader term** (2.3), **narrower term** (2.37)

NOTE Several different types of hierarchical relationship exist. For a further explanation, see 10.2.

2.24

homograph

one of two or more words that are written in the same way, but have different meanings

EXAMPLES

In English: The word "bank" could refer to a financial institution or the side of a river.
In French: The word "avocat" could refer to a lawyer or to a fruit.

NOTE Homographs are sometimes referred to as homonyms, although the latter term applies more broadly, as it also includes pairs of terms such as "weights" and "waits" in English or "mer" and "mère" in French, which sound the same although they are spelt differently.

2.25

identifier

set of symbols, usually alphanumeric, designating a **concept** (2.11) or a **term** (2.61) or another entity for purposes of unique identification within a determined context or resource, especially in a computer system or network

NOTE A notation is sometimes used as an identifier.

2.26

index term

term (2.61) assigned to a **document** (2.15) in the process of **indexing** (2.27)

NOTE Sometimes index terms are referred to as indexing terms, as keywords or as tags, but the latter terms have other meanings too. Preferred terms from a thesaurus are very often used as index terms.

2.27

indexing

intellectual analysis of the subject matter of a **document** (2.15) to identify the **concepts** (2.11) represented in it, and allocation of the corresponding **index terms** (2.26) to allow the information to be retrieved

NOTE The term "subject indexing" is often used for this concept, but as ISO 25964 does not deal with the indexing of other elements such as authors or dates, "indexing" is sufficient. Indexing can be carried out by human users or by automated agents.

2.28

information retrieval

all the techniques and processes used to identify **documents** (2.15) relevant to an information need, from a collection or network of information resources

NOTE Selection and inclusion of items in the collection are included in this definition; likewise browsing and other forms of information seeking.

2.29

interoperability

ability of two or more systems or components to exchange information and to use the information that has been exchanged

NOTE Vocabularies can support interoperability by including relations to other vocabularies, by presenting data in standard formats and by using systems that support common computer protocols.

2.30

loan term

term (2.61) borrowed from another language that has become accepted in the borrowing language

EXAMPLES

"glasnost" is a Russian term that has become accepted in English "gourmet" is a French term that has become accepted in English
--

2.31

markup

annotations or other type of encoding embedded in text, in conformity with a **markup language** (2.32)

2.32

markup language

set of encoding conventions that can be used to provide instructions for the interpretation of a text, by the use of annotations embedded in the text itself

NOTE The interpretation often concerns issues such as content, structure or rendering of the text. Widely used examples include HTML (Hypertext Markup Language)^[59], which is largely concerned with presentation, and XML (Extensible Markup Language) ^{[63][64][65][66]}, which addresses the structure of text.

2.33

metadata

data that identify attributes of a **document** (2.15) typically used to support functions such as location, discovery, documentation, evaluation and/or selection

NOTE Preferred terms or notations selected during the indexing process are commonly applied as metadata values.

2.34

monohierarchical structure

hierarchical arrangement of **concepts** (2.11), in a **thesaurus** (2.62) or **classification scheme** (2.6), in which each **concept** (2.11) can have only one broader **concept** (2.11) at the level immediately above

cf. **polyhierarchical structure** (2.42)

EXAMPLE In a monohierarchical structure, the concept of pianos cannot be listed under keyboard instruments as well as under stringed instruments; a choice has to be made of one of these concepts to determine its placing.

2.35

multilingual thesaurus

thesaurus (2.62) in which **terms** (2.61) and relational structures are available in two or more natural languages

2.36

multi-word term

term (2.61) consisting of more than one word

cf. **compound term** (2.9)

EXAMPLE

cost benefit analysis

2.37

narrower term

preferred term (2.45) representing a **concept** (2.11) that is narrower than the one in question

NOTE The scope of the narrower concept falls completely within the scope of the broader concept. The relationship between the two is commonly indicated with the tag NT. For more explanation see 10.2.1.

2.38

node label

label inserted into a hierarchical or classified display to show how the **terms** (2.61) have been arranged

NOTE A node label is neither a preferred term nor a non-preferred term. It contains one of two different types of information:

- a) the name of a facet to which following terms belong; or
- b) the attribute or characteristic of division by which an array of sibling concepts has been sorted or grouped.

See examples in Clause 11.

2.39

non-preferred term

non-descriptor

term (2.61) that is not assigned to **documents** (2.15) but is provided as an entry point in a **thesaurus** (2.62) or index

cf. **entry term** (2.16)

EXAMPLE

<i>hounds</i> USE dogs

NOTE In this example, "hounds" is a non-preferred term, while "dogs" is the preferred term that should be used in its place.

2.40

notation

class code

class number

classmark

set of symbols representing a **concept** (2.11) in a **structured vocabulary** (2.56), especially a **classification scheme** (2.6)

EXAMPLES

Notation	Source vocabulary	Concept
07.04.4	ILO Thesaurus	fishery policy and development
622.342 2	Dewey Decimal Classification	gold mining
373.3.016:51	Universal Decimal Classification	mathematics curriculum in primary schools
SBS XEJ B	Bliss Bibliographic Classification	endangered species law
H40-H42	International Statistical Classification of Diseases and Related Health Problems	glaucoma

NOTE Notation is sometimes used to sort and/or locate concepts in a predetermined systematic order and, optionally, to display how the components of complex concepts have been structured and grouped. A notation can provide the link between alphabetical and systematic lists in a thesaurus. In the context of classification schemes, "concepts" are often known as "subjects", especially when they are complex, as in the examples above.

2.41

paradigmatic relationship

a priori relationship

relationship between **concepts** (2.11) that is inherent in the **concepts** (2.11) themselves

NOTE Such relationships are shown in a structured vocabulary, independently of any indexed document. For a more complete discussion of paradigmatic and syntagmatic relationships, see 4.3.

2.42

polyhierarchical structure

hierarchical arrangement of **concepts** (2.11), in a **thesaurus** (2.62) or **classification scheme** (2.6), in which each **concept** (2.11) can have more than one broader **concept** (2.11)

cf. **monohierarchical structure** (2.34)

EXAMPLE

In a polyhierarchical structure, organs (musical instruments) could be listed under keyboard instruments as well as under wind instruments.

NOTE In a polyhierarchical structure, a single concept can occur in more than one place in the hierarchical structure of the thesaurus. Its attributes and relationships, and specifically its narrower and related terms, are the same wherever it occurs.

2.43

post-coordination

combination of **preferred terms** (2.45) of a **controlled vocabulary** (2.12) at the time of searching

cf. **pre-coordination** (2.44)

EXAMPLE

The post-coordinated search expression "microwaves AND radiation" can be used to retrieve documents on microwave radiation, when these have been indexed under the separate terms "microwaves" and "radiation" rather than a compound term.

2.44

pre-coordination

combination of **concepts** (2.11), classes or **terms** (2.61) of a **controlled vocabulary** (2.12) at the time of its construction or at the time of using it for **indexing** (2.27) or **classification** (2.5)

cf. **post-coordination** (2.43)

EXAMPLE 1

The class "general theory", when placed within the broader class "music", refers only to the pre-coordinated subject "theory of music" and not to theory in general.

EXAMPLE 2

The pre-coordinated string "cardboard – recycling" might appear in a subject heading scheme or, if not enumerated there, might be synthesized by an indexer when needed for a particular document.

2.45

preferred term

descriptor

term (2.61) used to represent a **concept** (2.11) when **indexing** (2.27)

cf. **non-preferred term** (2.39)

NOTE A preferred term is usually a noun or noun phrase.

2.46

protocol

convention that defines the syntax, semantics and synchronization of the communication process between two computers in order to enable a particular service

2.47

quasi-synonym

near-synonym

one of two or more **terms** (2.61) whose meanings are generally regarded as different in ordinary usage but which may be treated as labels for the same **concept** (2.11), in a given **controlled vocabulary** (2.12)

EXAMPLES

diseases, disorders
earthquakes, earth tremors

2.48

related term

preferred term (2.45) representing a **concept** (2.11) that has an **associative relationship** (2.2) with the one in question

NOTE The relationship between related terms is commonly indicated with the tag RT. For a further explanation, see 10.3.

2.49

schedule

terms (2.61), **notations** (2.40), captions, cross-references and **scope notes** (2.50) set out to exhibit the content and structure of a **structured vocabulary** (2.56)

2.50

scope note

note that defines or clarifies the semantic boundaries of a **concept** (2.11) as it is used in the **structured vocabulary** (2.56)

NOTE A term used to label a concept can have several meanings in normal usage. A scope note is used to restrict the concept to only one of those meanings, and where necessary refers to other concepts that are included or excluded from the scope of the concept being clarified.

2.51

search term

term (2.61) forming all or part of a search query

NOTE In the context of ISO 25964, search terms are usually drawn from a controlled vocabulary.

2.52

sibling concept

one of two or more **concepts** (2.11) with the same immediate broader **concept** (2.11), each of these being represented by a **preferred term** (2.45)

EXAMPLE

In the following, outerwear and underwear are preferred terms representing sibling concepts in the same array:

clothing
 outerwear
 overcoats
 underwear

2.53

sibling term

one of two or more **preferred terms** (2.45) with the same immediate **broader term** (2.3)

EXAMPLE

In the following, chairs and tables are sibling terms in the same array, while no siblings are shown for “furniture”, “armchairs” or “dining tables”:

furniture
 chairs
 armchairs
 tables
 dining tables

2.54

source language

language serving as a starting point in translation or in a search for **term** (2.61) equivalents

2.55

specificity

capability of a **structured vocabulary** (2.56) to express a subject in depth and in detail

NOTE For a further explanation, see the discussion of specificity in 8.4 and other places.

2.56

structured vocabulary

organized set of **terms** (2.61), headings or codes representing **concepts** (2.11) and their inter-relationships, which can be used to support **information retrieval** (2.28)

NOTE A structured vocabulary can also be used for other purposes. In the context of information retrieval, the vocabulary needs to be accompanied by rules for how to apply the terms. Various types of structured vocabulary will be addressed in ISO 25964-2, including classification schemes, subject heading schemes, etc.

2.57

subject heading scheme

subject heading language

subject heading list

SHL

structured vocabulary (2.56) comprising **terms** (2.61) available for subject **indexing** (2.27), plus rules for combining them into pre-coordinated strings of **terms** (2.61) where necessary

2.58

synonym

one of two or more **terms** (2.61) denoting the same **concept** (2.11)

EXAMPLES

In English: guarantees, warranties heart attack, myocardial infarction HIV, human immunodeficiency virus
In French: schiste, phyllade VIH, virus de l'immunodéficience humaine crise cardiaque, infarctus du myocarde

NOTE Abbreviations and their full forms can be treated as synonyms.

2.59

syntagmatic relationship

a posteriori relationship

relationship between **concepts** (2.11) that exists only because they occur together in a **document** (2.15) being indexed

NOTE Such relationships are not generally valid in contexts other than the document being indexed, and therefore they do not form part of the structure of a thesaurus. For a more complete discussion of syntagmatic and paradigmatic relationships, see 4.3.

2.60

target language

language providing a translation or an equivalent for a **term** (2.61) existing in a **source language** (2.54)

2.61

term

word or phrase used to label a **concept** (2.11)

EXAMPLES

schools school uniform costs of schooling teaching

NOTE Thesaurus terms can be either preferred terms or non-preferred terms.

2.62

thesaurus

controlled (2.12) and **structured vocabulary** (2.56) in which **concepts** (2.11) are represented by **terms** (2.61), organized so that relationships between **concepts** (2.11) are made explicit, and **preferred terms** (2.45) are accompanied by lead-in entries for **synonyms** (2.58) or **quasi-synonyms** (2.47)

NOTE The purpose of a thesaurus is to guide both the indexer and the searcher to select the same preferred term or combination of preferred terms to represent a given subject. For this reason a thesaurus is optimized for human navigability and terminological coverage of a domain.

2.63

top term

preferred term (2.45) representing a **concept** (2.11) that has no broader **concept** (2.11) in the **thesaurus** (2.62)

2.64

vocabulary control

management of a vocabulary in order to disambiguate and constrain the form of the **terms** (2.61) and limit the number of **concepts** (2.11) and **terms** (2.61) available for **indexing** (2.27)

NOTE Control is achieved by distinguishing between homographs, so that each one has just one meaning, and by picking out from a set of synonyms or quasi-synonyms, the one which is to be preferred for use in indexing. The purpose of these restrictions is to increase the likelihood of indexers and searchers choosing the same term to label a particular concept.

3 Symbols, abbreviated terms and other conventions

3.1 Table 1 shows the abbreviated terms that are used in English language examples of thesaurus records as tags to prefix terms and notes. Each tag indicates the relationship or function of the term or note that follows.

Table 1 — Symbols and abbreviations

Description	Symbol	Tag	Meaning
<i>Descriptive elements</i>		SN	Scope note
		DEF	Definition
		HN	History note
<i>Codes</i>		SC	Subject category; a code or notation applied to a group of concepts relevant to a particular subject
		CC	Concept code or notation
<i>Relationships</i>	→	USE	Use; the term that follows the tag is the preferred term that should be used in place of the non-preferred term that precedes the tag
	=	UF	Use for or Used for; the term that follows the tag is a non-preferred term for which the preferred term preceding the tag should be used instead
		USE...+	The two or more preferred terms following the tags should be used together to represent the indicated concept
		UF+	The non-preferred term that follows should be represented by a combination of preferred terms including the preferred term that precedes the tag
		TT	Top term; the preferred term that follows the tag represents the broadest concept in a hierarchy to which the specific concept belongs
	<	BT	Broader term; the term that follows the tag represents a concept having a wider meaning
		BTG	Broader term (generic)
		BTI	Broader term (instantial)
	—<	BTP	Broader term (partitive)
	>	NT	Narrower term; the term that follows the tag refers to a concept with a more specific meaning
		NTG	Narrower term (generic)
		NTI	Narrower term (instantial)
	>—	NTP	Narrower term (partitive)
	—	RT	Related term; the term that follows the tag is associated, but is not a synonym, a quasi-synonym, a broader term or a narrower term

Table 2 — English language tags and their equivalents in other languages

Tags in English	Tags in French	Tags in German	Tags in Danish	Tags in Finnish	Tags in Norwegian	Tags in Swedish	Tags in Spanish	Tags in Chinese	Tags in Russian
SN Scope note	NE Note explicative ^b <i>Note d'emploi</i>	H Erläuterung (Hinweis)	SN Forklarende note	Huomautus ^c	NO Note	SN Anmärkning	NA Nota de aplicación <i>Nota de alcance</i>	J 解	лп Лексическое примечание
USE Use	EM Employer	BS Benutze	USE Brug	KÄ käytä	BRUK Bruk	USE Använd	USE Use	Y 用	см Смотри
UF Use for <i>Used for</i> ^a	EP Employé pour <i>Employer pour</i>	BF Benutzt für <i>Benutze für</i>	UF Brugt for	KT korvattu termi	BF Brukt for	UF Används istället för	UP Usado por <i>Use por</i>	D 代	с Синоним
BT Broader term	TG Terme générique	OB Oberbegriff	BT Overordnet term	LT laajempi termi	OT Overordnet term	BT Överordnad term	TG Término genérico	S 属	в Выше
NT Narrower term	TS Terme spécifique	UB Unterbegriff	NT Underordnet term	ST suppeampi termi	UT Underordnet term	NT Underordnad term	TE Término específico	F 分	н Ниже
RT Related term	TA Terme associé	VB Verwandter Begriff	RT Relateret term	RT rinnakkaistermi	SO Se også	RT Relaterad term	TR Término relacionado	C 参	а Ассоциация
BTG Broader term (generic)	TGG Terme générique (générique)	OA Oberbegriff (Abstraktionsrelation)	BTG Overordnet term (generic)						вр Выше—род
BTP Broader term (partitive)	TGP Terme générique (partitif)	SP Verbandsbegriff (Bestandsrelation)	BTP Overordnet term (partitiv)						вц Выше—целое
BTI Broader term (instantial)	TGI Terme générique (instance)		BTI Overordnet term (instans)						
NTG Narrower term (generic)	TSG Terme spécifique (générique)	UA Unterbegriff (Abstraktionsrelation)	NTG Underordnet term (generic)						нв Ниже—вид
NTP Narrower term (partitive)	TSP Terme spécifique (partitif)	TP Teilbegriff (Bestandsrelation)	NTP Underordnet term (partitiv)						нч Ниже—часть
NTI Narrower term (instantial)	TSI Terme spécifique (instance)		NTI Underordnet term (instans)						

NOTE The selection of languages in Table 2 is open-ended. Future editions of this part of ISO 25964 could include additional tags.

^a Lines in italics show variant tags and/or expansions.^b French AFNOR Standard (AFNOR Z47-100) proposed NA: Note d'application^c In the Finnish standard "Scope note" is translated "Selitys" at the moment, but "Huomautus" would be a better translation. There is no associated tag.

3.2 Thesauri in Dutch use the same tags as in English, without translation.

3.3 The abbreviations listed as tags in Table 2 have acquired status as generally recognized conventions, and they occur in many published thesauri. They have mnemonic value, but apply only in particular languages. If a more neutral system of abbreviation is required, the language-independent symbols in the left-hand column of Table 1 are recommended.

3.4 The following conventions are also used in examples throughout this part of ISO 25964. Note that these conventions are not required of all thesauri.

- a) Preferred terms are printed in the case they would normally have in full text. Typically that means lower case throughout, with the exception of upper case for abbreviations and acronyms, and initial capital letters for proper names.

EXAMPLES

cars
animals

- b) Non-preferred terms are printed in italics.

EXAMPLES

cars	<i>automobiles</i>
UF <i>automobiles</i>	USE cars
animals	<i>fauna</i>
UF <i>fauna</i>	USE animals
World Health Organization	<i>WHO</i>
UF <i>WHO</i>	USE World Health Organization

4 Thesaurus overview and objectives

4.1 Overall objective

The traditional aim of a thesaurus is to guide the indexer and the searcher to choose the same term for the same concept. In order to achieve this, a thesaurus should first list all the concepts that might be useful for retrieval purposes in a given domain. The concepts are represented by terms, and for each concept, one of the possible representations is selected as the preferred term (see 4.2 and 6.6). Secondly, a thesaurus should present the preferred terms in such a way that people will easily identify the one(s) they need. This is achieved by establishing relationships (see Clauses 8 to 10) between terms — and/or between concepts — and using the relationships to present the terms in a structured display.

When implemented in a search system or indexing system, the terms and relationships in the thesaurus may be exploited in numerous ways, including:

- search expansion;
- suggestion of alternative search terms;
- support for clustering or other means of refining a search;
- identification of common spelling mistakes;
- support of automatic indexing.

The interplay between terms, concepts and relationships is illustrated in the data model in Clause 15. For reasons of practicality this part of ISO 25964 refers frequently to "terms". But it should never be forgotten that the purpose of manipulating terms is to support retrieval of the underlying concepts.

4.2 Vocabulary control and its purpose

Establishing an appropriate preferred term to represent a particular concept is not always straightforward because a concept can often be expressed in more than one way. Furthermore, in ordinary discourse one term can have more than one meaning. Vocabulary control is therefore essential, and thesauri are used to achieve this by the following two principal means.

- a) Concepts and terms are deliberately restricted in scope to selected meanings. Unlike the terms in a dictionary, which are often accompanied by a number of different definitions reflecting common usage, each term in a thesaurus is generally restricted to whichever single meaning serves the needs of a retrieval system most effectively. The structure of a thesaurus, notably its display of hierarchical relationships, frequently indicates the intended meaning of a term. If this technique is not sufficiently explicit, a scope note should accompany the term. This note should state the chosen meaning, and it may also indicate other meanings that are recognized in natural language but which have been deliberately excluded for indexing purposes.
- b) When the same concept can be expressed by two or more synonyms or quasi-synonyms in the same language, one of these terms is usually selected as the preferred term, which is then used consistently in indexing or simply as the first or only term representing the concept. Reference to the preferred term should be made from any synonym that might also function as a user's access point.

One consequence of using the measures in a) and b) for vocabulary control is that the resultant language might not correspond to a user's preferences. The thesaurus has an important role in mediating between terms used in discourse and those that function effectively for information retrieval. To achieve the retrieval benefits, users need to accept a degree of artificiality in the controlled vocabulary (although in some systems this difficulty can be overcome by automatic substitution of the user's term with the preferred form).

So that a thesaurus can function effectively in a multilingual context, the concepts included need to be represented in all of the languages present, enabling speakers of these languages to have access to them. If the thesaurus is "symmetrical" as described in this part of ISO 25964 and modelled in Clause 15, each concept should have a preferred term in each language, and the scope should be the same in all the languages. This constraint sometimes adds to the artificiality of the vocabulary. Alternative approaches (for example the use of mappings between vocabularies) will be described in ISO 25964-2.

4.3 Paradigmatic versus syntagmatic relationships

The purpose of establishing and displaying relationships is likewise to guide users (or intelligent agents acting on their behalf) to choose the most appropriate term(s) for expressing a given concept. This is achieved by presenting terms that a user might consider instead of, or as well as, the term entered first. Means for displaying relationships are described in Clause 12.

Two kinds of relationship are commonly encountered between indexing terms in any one language.

- a) Relationships that exist only because the concepts occur together in the context of a particular document are called syntagmatic relationships. When two or more indexing terms are assigned to the same document, this act effectively records a syntagmatic relationship between them. In a thesaurus, however, inclusion of syntagmatic relationships is not recommended.

EXAMPLE 1

An indexer dealing with a work on "Computers in banks in Amsterdam" might identify three concepts represented respectively by the terms: "banks (financial institutions)", "computers" and "Amsterdam". He assigns these to the document. In a post-coordinate system the relationship between these terms is not explicitly indicated in metadata, but the document would nonetheless be retrieved if any or all of these terms were used as retrieval keys. In a pre-coordinated index an entry can be provided for the combination of three terms, with a reference to the location where the corresponding document can be found. The concepts in this example are not normally associated according to common frames of reference, and their interrelationships can, therefore, be regarded as document-dependent.

- b) Relationships that are valid in almost all contexts, especially when they are inherent in the concepts that the terms represent, are known as paradigmatic relationships. It is useful to show paradigmatic relationships between thesaurus concepts, as they will frequently guide users to concepts closely related to the terms they first thought of.

EXAMPLE 2

Using concepts from the previous example, "banks (financial institutions)" has an inherent relationship with the broader concept "financial institutions"; "computers" has a strong association with "data processing", independent of the subject of the document being indexed; and "Amsterdam" is inseparably linked with "Netherlands". A user looking up any of these linked terms might be interested in information indexed with the former terms. These paradigmatic relationships are independent of any particular document. They are generally recognized and could be established through reference to standard works, such as dictionaries and encyclopaedias.

The distinction between these two kinds of relationships can be displayed as shown in Figure 1.

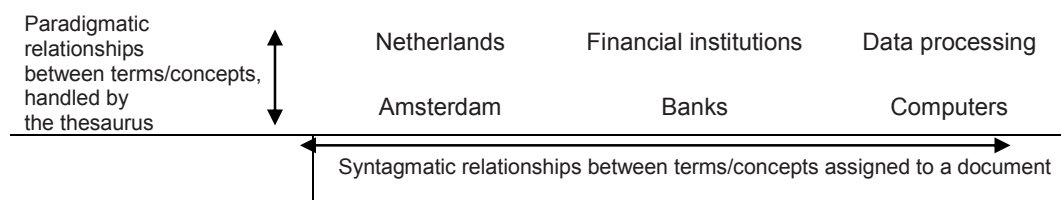


Figure 1 — Paradigmatic and syntagmatic relationships

4.4 Types of paradigmatic relationship

Three types of paradigmatic relationship between terms, a), or between concepts, b) and c), should be established and clearly distinguished:

- the equivalence relationship as it applies in both monolingual and multilingual situations (see Clauses 8 and 9 respectively);
- the hierarchical relationship (see 10.2);
- the associative relationship (see 10.3).

Further subdivisions of each of these classes are described in Clauses 8 to 10. Each of these relationships should be established reciprocally, and this should be indicated by a system of tags, symbols or abbreviations used to represent relationships in a thesaurus. The conventional tags described in Clause 3 are used in the examples that follow.

5 Concepts and their scope in a thesaurus

5.1 Conceptual basis

5.1.1 The prime application for a thesaurus is in information retrieval, where the aim is to search for concepts. As explained in 4.1 and 4.2, the concepts are represented by terms. Each term included in a thesaurus should represent a single concept (or unit of thought). Concepts can range from the simple (e.g. "cats") to the very complex (e.g. "racial discrimination among ethnic minorities"). Compound terms or phrases are generally needed to express the more complex concepts. See Clause 7 for a more complete discussion of these.

5.1.2 The concepts included in a thesaurus may be considered as belonging to mutually exclusive categories based on shared characteristics. The following examples are indicative but do not list all the possible categories.

a) Things and their physical parts.

EXAMPLE SET 1

birds birth certificates limbs microforms monuments mountain regions

NOTE Physical parts are also things, and can have their own parts.

b) Materials.

EXAMPLE SET 2

adhesives rubber titanium

c) Activities or processes.

EXAMPLE SET 3

dressmaking fertilization glaciation land management

d) Events or occurrences.

EXAMPLE SET 4

birthdays civil wars revolutions
--

- e) Properties of persons, things, materials or actions.

EXAMPLE SET 5

consciousness
elasticity
personality
speed

- f) Disciplines or subject fields.

EXAMPLE SET 6

archaeology
organic chemistry
theology

- g) Units of measurement.

EXAMPLE SET 7

hertz
kilometres

- h) Types of people and organizations.

EXAMPLE SET 8

charities
children
international nongovernmental organizations
nations
poets
visually impaired people

5.1.3 Unique entities expressed as proper names may also be included in a thesaurus, where they should be used only for records that are *about* the person, document or object, etc. When the entity represented by the proper name is not the subject of a document, but has a different relationship, such as being its creator or title, the proper name should not be recorded together with subject terms, but elsewhere.

Proper names also belong to categories such as those listed in 5.1.2.

- a) Places.

EXAMPLE SET 1

Australia
Milky Way
South Kensington
Sri Lanka

- b) Specific objects, topographical features, and other entities.

EXAMPLE SET 2

Magna Carta
Mona Lisa
Nelson's Column
Romeo and Juliet
Skylab

- c) Individuals, organizational posts, and corporate bodies.

EXAMPLE SET 3

Burns, Robert
United Nations Secretary-General
World Health Organization

5.2 Scope notes

The intended scope of a concept in the thesaurus is not always the most common meaning associated with its preferred term (or any one of its non-preferred terms) in ordinary discourse. In the thesaurus, context is usually provided by the hierarchy of broader and narrower concepts linked to the concept in question and this helps to elucidate the intended scope. Often it is possible to formulate the preferred term of the concept in an unambiguous way that conveys the intended scope to any user (see 6.2.2). However, in cases where these measures are not appropriate or sufficient, or where additional information would help to clarify the meaning and make the usage more consistent, an explicit scope note should be used.

A scope note should be used to clarify the boundaries of a concept, especially when the meaning of the preferred term in ordinary discourse can be interpreted too broadly or too narrowly, or to distinguish between preferred terms that have overlapping meanings in natural language. It can also be used to provide other advice on term usage to either the indexer or the searcher. A scope note need not be a full definition but should clarify the intended use of a term within the thesaurus. Other types of note may be conveyed separately; see 6.2.3 for definitions and 6.2.4 for history notes.

Unlike the qualifiers considered in 6.2.2, a scope note is not regarded as forming part of the term to which it is attached.

EXAMPLE 1

microwave frequencies
SN 1 GHz to 300 GHz

EXAMPLE 2

illuminations
SN Includes both the ornamental decoration and the illustrations in manuscripts, as well as in some early printed books, if done by hand

5.3 Reciprocal scope notes

When reference is made to other concepts in a scope note, a reciprocal scope note should generally be provided for each concept mentioned.

EXAMPLE

In English: food products SN Use only for products intended for human consumption. For products for animals, see petfoods or feeds. petfoods SN Food products for animals maintained as domestic pets. For products for human consumption, see food products. For products intended for non-domestic animals, see feeds. feeds SN Products intended for non-domestic animals. For products for domestic pets, see petfoods. For products for human consumption, see food products.	In French: statistique NE Désigne la science statistique; pour les données statistiques employer "statistiques" statistiques NE Désigne les données statistiques; pour la science statistique, employer "statistique"
---	--

Even where the scope of only one of the concepts requires clarification, it is useful to make an editorial note (see 15.2.14) in the record for any additional concept, of where it has been cited. The purpose of the reciprocal reference is to ensure that when a change is made to one of the concepts, or it is deleted, the effect on the other concept is considered.

6 Thesaurus terms

6.1 Form of terms

The terms selected to represent concepts may be single-word terms or multi-word terms (see Clause 7).

6.2 Clarification and disambiguation of thesaurus terms

6.2.1 General

In ordinary discourse, a single term might have more than one meaning, and the most common meaning is not always the one needed to represent the concept wanted in the thesaurus. If the context provided by the hierarchy of broader and narrower concepts linked to the concept in question is insufficient to elucidate the intended scope, additional measures should be applied. While all thesaurus terms should be expressed as unambiguously as possible, it is especially important to formulate the preferred term of a given concept in such a way that it conveys the intended scope to any user. For example, the multi-meaning term "depression" could be reformulated as "economic depression" or "meteorological depression", as appropriate. Alternatively a qualifier (see 6.2.2) may be used. In cases where these measures are not appropriate or sufficient, or where additional information would help to clarify the meaning and make the usage more consistent, an explicit scope note should be used (see 5.2).

6.2.2 Homographs and qualifiers

Homographs (sometimes referred to by the broader term "homonyms") are words with the same spelling but different meanings.

EXAMPLE 1

cranes (This term can refer either to birds or to lifting equipment.)

When homographs are needed as thesaurus terms, the meaning of each term should be clarified and the traditional way to do this is by adding to it a qualifier in parentheses. The qualifier should be as brief as possible, ideally consisting of one word. Often a broader term, the qualifier should indicate the context or subject area to which the concept belongs. It forms part of the term and does not serve as a scope note (see 5.2).

EXAMPLE 2

cranes (birds)
cranes (lifting equipment)

A qualifier should be added to each homographic term, even when one of its senses is more common than the other in the domain of the thesaurus. For example in English, "beams (structures)" is an appropriate preferred term in an engineering thesaurus that also includes "beams (radiation)". For example in French, "eau (boisson)" is an appropriate preferred term in an environmental thesaurus that also includes "eau (environnement)".

If only one of the homographic terms is used in a specialized thesaurus, and its meaning is obvious to the users of the thesaurus, the qualifier may be left out. However, consideration should be given to the possibility that the thesaurus scope may be expanded in the future, or that interoperability with other vocabularies covering different scopes might be needed.

Since the qualifiers make the terms slightly cumbersome to apply, and since some electronic systems have difficulties in applying them, their use (especially in preferred terms) should be avoided if another means of resolving ambiguity can be found. For this reason, a multi-word term should be preferred to a single-word term with a qualifier, as long as the compound form occurs in natural language.

EXAMPLE 3

For example, "industrial plants" should be preferred to "plants (facilities)". The latter expression, however, might be entered as a non-preferred term.

Parenthetical qualifiers should not be used to provide a term in inverted form.

EXAMPLE 4

For example the inverted forms "cookery (fish)" and "pens (fountain)" are incorrect and the direct forms "fish cookery" and "fountain pens" should be used instead. "Fountain" is used in the latter example to indicate a type of pen rather than to disambiguate the word "pen". (See also 7.7 on the order of words in multi-word terms). Appropriate uses of qualifiers with the term "pens" in English are: "pens (enclosures)" and "pens (writing implements)". Appropriate uses of qualifiers with the term "audition" in French are: "audition (physiologie)" and "audition (communication)".

In the case of acronyms and abbreviations, which are often homographs, the qualifier should be the full form of the term (see also 6.6.8). Qualifiers for acronyms and abbreviations should only be used when their meaning is considered to be ambiguous to the users of the thesaurus, i.e. when they have another known meaning either in the domain covered by the thesaurus or in common language.

6.2.3 Definitions

A full definition is not usually required to clarify the way in which a preferred term should be used. However, if a definition is required for another reason, a separate note field should be established for the definitions so that they do not become confused with any scope notes. The source of each definition should be recorded alongside the definition itself.

EXAMPLE

chiaroscuro DEF The style of pictorial art in which only the light and shade are represented (OED)
--

NOTE "OED" refers to the Oxford English Dictionary (www.oed.com). Any such acronyms in thesaurus definitions or scope notes should be explained in the introduction to the thesaurus (see 13.4).

6.2.4 History notes

When a thesaurus is updated, changes in terms can affect the retrievability of a concept. In such cases a history note may be used.

A history note may apply to a single preferred or non-preferred term, or to a concept. It should be used when a new preferred term is added to the thesaurus or a change is made to an existing term that affects the concept's scope in different periods of application. While it is possible to include such information in the scope note, a separate history note is preferable. A history note may record the date of introduction, or it may give more complex advice on how to search for the concept in earlier or later times.

EXAMPLE 1

microwave ovens HN Concept introduced 1985
--

EXAMPLE 2

notebook computers HN Term introduced 1999; prior to that use "laptop computers"
--

6.3 Grammatical form of terms

6.3.1 Nouns and noun phrases

A thesaurus term should preferably consist of a noun or a noun phrase. This includes gerunds, which are verbal nouns (see 6.3.4). In the English language (and in others, such as German, French and other romance languages) noun phrases occur in two forms:

- a) adjectival phrases (including some where a noun or possessive functions as adjective)

EXAMPLE SET 1

Burkitt's lymphoma
cold fusion
environmental luminosity
stone walls
tropical diseases

- b) prepositional phrases

EXAMPLE SET 2

accessories after the fact
hospitals for children
prisoners of war

As prepositions can unnecessarily add to the length and clumsiness of a term, they should be avoided if possible, especially in preferred terms. For example, use "carbohydrate metabolism" rather than "metabolism of carbohydrates". (However, the latter term could still be admitted as a non-preferred term.)

In some other languages such as Chinese, prepositional phrases do not occur, but other types of noun phrase are available and should be used.

6.3.2 Adjectives

Adjectives can be useful as components of noun phrases, but if used standing alone they are a potential cause of problems in retrieval. A search for an article about "the use of red lights as warning signals for low bridges", if indexed with adjectival terms such as "red" and "low", could result in the retrieval of information about low lights and/or red bridges.

For this reason the use of adjectives as thesaurus terms should be avoided. The same caution applies even if the adjective is converted to the corresponding noun, e.g. "red" to "redness".

Exceptions may be made where there are strong reasons, for example in a thesaurus for indexing reports of medical symptoms or for images or other non-text collections, where adjectives describing the appearance might be very important.

EXAMPLES

dark red
large
oval
smooth

6.3.3 Adverbs

Adverbs such as "very" or "highly" should not be used as thesaurus terms. A phrase beginning with an adverb should not be accepted as a thesaurus term unless it has acquired a special meaning.

EXAMPLES

very high frequency very large scale integration

6.3.4 Verbs

In the English language verbs expressed as infinitives or participles should not be used alone as thesaurus terms, unless they are expressed as gerunds (for example, "weaving", "broadcasting"). Activities should be represented by nouns or verbal nouns. For some other languages, infinitives may be acceptable and this recommendation may be adapted to respect commonly accepted conventions for indexes.

EXAMPLES

cookery or cooking (not "cook", "to cook", "cooked", etc.) distillation (not "distil") swimming (not "swim")
--

6.3.5 Initial articles**6.3.5.1 Omission**

The use of initial articles in thesaurus terms should generally be avoided. If necessary, a parenthetical qualifier should be used.

EXAMPLES

arts	<i>rather than</i>	the arts
state (political entity)	<i>rather than</i>	the state

6.3.5.2 Retention

If the initial article is an integral part of a proper name and needs to be searchable, it should be included in the thesaurus term in direct order. Otherwise, the article should be omitted or the term inverted. If the omission causes ambiguity, a qualifier should be added. Whether or not the article is considered an integral part of the name depends on the language and the context. In the following examples, English is the assumed language of the thesaurus.

EXAMPLE 1

El Niño Le Havre Los Angeles Needles (Isle of Wight) The Who (rock music group)

Where a term might be sought with or without the article, a reference should be made from the non-preferred form.

EXAMPLE 2

<i>Salvador (country)</i>	USE	El Salvador
<i>The Needles</i>	USE	Needles (Isle of Wight)
<i>Who, The</i>	USE	The Who (rock music group)

6.4 Capitalization, punctuation and special characters

6.4.1 Capitalization

A consistent style should be used for the presentation of thesaurus terms. This part of ISO 25964 uses lower case throughout, except where capitals are required in proper names. As computers are no longer restricted to using upper case only, either lower case or lower case with initial capitals should be used.

Exceptions may be made for abbreviations, acronyms, proper names or terms that are conventionally written in a special style. In these cases, the style that is most widely accepted among the anticipated users of the thesaurus should be used.

EXAMPLES

British Airways Plc
ActiveX
DNA
NPK fertilizers
pH
photocopies

6.4.2 Non-alphabetic characters

The use of punctuation marks, diacritics and other special characters can cause problems for entering terms, as well as sorting and processing in search applications. Apostrophes, parentheses, hyphens and numerical characters can complicate a search expression. The use of such characters should be minimized, especially in preferred terms.

EXAMPLE SET 1 (of avoidance of special characters)

In English:		
beta rays	<i>Not</i>	β-rays
databases	<i>Not</i>	data-bases
nonfiction	<i>Not</i>	non-fiction
research and development	<i>Not</i>	research & development
In French:		
rayon bêta	<i>au lieu de</i>	rayon β
recherche et développement	<i>au lieu de</i>	recherche & développement

However, as their use cannot be avoided completely, they should be retained wherever the terminology would otherwise be ambiguous, ungrammatical or unacceptable to the user community of the thesaurus, particularly in abbreviations, chemical names, proper names and trademarks, or if these terms belong to a standardized vocabulary in the domain. For thesauri in languages other than English, diacritics are often essential.

EXAMPLE SET 2 (where special characters cannot be avoided)

In English: 2,4-D bis(tributyltin) oxide Boyle's law Burkitt's lymphoma X-rays
In French: 2,4-D oxyde de bis(tributylétain)

Hyphens and diacritics should also be retained in non-preferred terms, especially when they are used as commonly recognized spelling variants in the language of the thesaurus.

EXAMPLE SET 3 (of special character retention in non-preferred terms)

In English: non-fiction résumés	USE USE	nonfiction resumes
In French: contre-marche contre-plaqué	EM EM	contremarche contreplaqué

Where possible (but not in the case of names of chemicals, and others where they form an integral part of the term) the use of parentheses should be limited to qualifiers.

6.5 Singular or plural forms

6.5.1 Cultural and linguistic factors

Different conventions exist in different natural languages concerning the use of singulars or plurals. Indexers in some language communities, for example French and German, tend to use the singular form so that the user can approach and use the thesaurus in the same way he would approach and use a dictionary. In English and Spanish, however, it is usual to base the choice on whether a particular term is a count noun or a non-count noun. The latter convention helps to distinguish between a process such as "painting", which can only be expressed in the singular, and the product of the same process, in this case "paintings". The form of the terms in each language should be based on the conventions applied in this language. As a consequence of such practice, a multilingual thesaurus is liable to have entries in which a term in the singular in French or German has an English equivalent in the plural form.

EXAMPLE

fr:	maison
de:	Haus
en:	houses
es:	casas

6.5.2 Treatment of count nouns in English and Spanish

NOTE The guidelines in this subclause are known to apply to English and Spanish, but not to French and German. For any other language, the best approach can be established by investigating practice in widely used thesauri in that language.

Count nouns are names of countable entities that are subject to the question "How many?" but not "How much?". They should be expressed as plurals.

EXAMPLE SET 1 (of count nouns)

documents
penguins
political parties
windows

One exception is the names of parts of the body, which are usually expressed in the singular.

EXAMPLE SET 2 (of exceptions)

digestive system
eye
head

Another exception is the names of living organisms. Many species, such as *Escherichia coli* or *Euonymus fortunei*, do not have a common name and are known only by their scientific name, which by convention is expressed in Latin, in the singular. In the interest of consistency, when Latin names and names accepted as English coexist in one compilation, the singular may be applied throughout. Use of the singular is not mandatory, however, and thesaurus editors may prefer the plural, for example "zebras", "whales" or "daffodils". Once a convention has been chosen, it should be applied consistently.

When a controlled vocabulary is used for naming museum objects, the singular forms of terms are commonly used. In general, however, concepts represent categories of objects rather than individual objects, and the plural form is more appropriate. A single item might be named a "chair", but the *category* likely to be sought by an enquirer is "chairs". Using the plural form of the preferred term in these cases can add consistency so that the same thesaurus may be used for museum objects as for other types of information resources.

6.5.3 Treatment of non-count nouns in English and Spanish

NOTE The guidelines in this subclause are known to apply to English and Spanish, but not to French and German. For any other language, the best approach can be established by investigating practice in widely used thesauri in that language.

Non-count nouns are names of concepts such as materials or substances that are subject to the question "How much?" but not "How many?". They should usually be expressed as singulars.

EXAMPLE SET 1 (of non-count nouns)

cotton
aluminium
steam

However, if the community of users served by the thesaurus regards a given substance or material as a class with more than one member, the class should be expressed in the plural.

EXAMPLE SET 2 (of exceptions)

grasses poisons steels

The names of abstract phenomena, properties, systems of belief, activities and disciplines are often non-count nouns that should be expressed in their singular forms.

EXAMPLE SET 3 (additional non-count nouns)

Abstract phenomena:	personality; winter
Properties:	brittleness; opacity; solubility
Systems of belief:	Catholicism; Shintoism; communism
Activities or processes:	cutting; immigration; shrinkage
Disciplines:	astronomy; sociology

However, when an abstract concept is regarded as a class with more than one member, the term representing the class should be expressed in the plural.

EXAMPLE SET 4 (additional exceptions)

chemical reactions intelligence tests physical sciences

6.5.4 Coexistence of singular and plural

In any language, where the singular and plural forms of a term refer to different concepts both should be entered in the thesaurus. The distinction between them should be reinforced by adding a scope note and if possible a qualifying term or phrase.

EXAMPLE 1

In English: wood (material) woods (areas of woodland)
In French: statistique (science) statistiques (données)

The added qualifier becomes an integral part of the term; it does not constitute a scope note (see also 5.2).

Where singular and plural denote the same concept, but their spelling differs to such an extent that the terms would be separated by unrelated terms when filed alphabetically, a reference should be made from the non-preferred form.

EXAMPLE 2

In English:		
mouse	USE	mice
In French:		
yeux	EM	oeil

Where singular and plural denote the same concept, and their spelling differs very little, e.g. "apron" and "aprons", it is usual to provide an entry only for the preferred term. But the non-preferred form may be added as an aid when computer look-up is anticipated.

6.6 Selection of the preferred form

6.6.1 General

Sometimes more than one term in the same language is available to represent the same concept. When multiple terms are included corresponding to the same concept, one of these should be designated the preferred term and the others non-preferred terms (see Clause 8). The situation across languages is described in Clause 9.

Where there is a choice between synonymous forms of expression, the preferences of the community to be served by the thesaurus should be adopted (subject to the avoidance of ambiguity or language that might be offensive to some users).

6.6.2 Spellings

The most widely accepted spelling of words should be adopted for preferred terms. If variant spellings exist and are commonly recognized, each should be entered in the thesaurus as a non-preferred-term, and a reference should be made from the non-preferred term to the preferred term.

EXAMPLE 1

In English:		
<i>Roumania</i>	USE	Romania
<i>Rumania</i>	USE	Romania
In French:		
<i>Lithuanie</i>	EM	Lituanie

EXAMPLE 2

In English:		
non-fiction	USE	nonfiction
In French:		
co-voiturage	EM	covoiturage

Where possible, spelling should follow the practice of a well-established dictionary or glossary in the language of the thesaurus. If a choice between spellings is made for reasons of dialect (for example between American English and British English), the chosen source should be stated in the introduction to the thesaurus (see 13.4), and the choice should be adhered to consistently throughout.

However, proper names can cause legitimate inconsistencies in spelling practice, as they should be spelt according to the practice of their owners or recognized authorities such as the Virtual International Authority File (VIAF)¹. See Example set 3, which assumes British spelling for terms other than proper names as appropriate.

EXAMPLE SET 3

colour	UF	<i>color</i>
defence	UF	<i>defense</i>
Department of Defense	(a US government department)	
Ministry of Defence	(a UK government department)	

Misspelt words are not acceptable in preferred terms. However, some misspellings are so common that they can provide useful entry points, especially in electronic thesauri, and they may be used as or in non-preferred terms.

EXAMPLE 4

In English:	
abattoirs	
UF	<i>abatoirs</i>
	<i>abbatoirs</i>
	<i>abbattoirs</i>
In French:	
Mitterrand François	
EP	<i>Mitterand François</i>

Misspelt entry points should be provided only where:

- a) the misspelling is common enough to be worthwhile; and
- b) there is no danger that the misspelling relates to a term other than the preferred term indicated.

The entries should be marked in some way to clarify that the terms are misspelt. One option is to substitute a tag such as MS (meaning MisSpelling) for UF as the introductory tag.

When misspelt words are entered as non-preferred terms, a choice may be made to suppress the misspelt entries in printed or electronic displays, where they add nothing to a user's understanding of the concept, but retain them in electronic versions used to interpret search terms entered by users.

1) VIAF is a joint project of several national libraries (17 at the time this part of ISO 25964 was prepared, contributing over 13 million personal name authority records), hosted by OCLC and available at <http://viaf.org/>.

6.6.3 Loan terms and translations of loan terms

Loan terms may be used as preferred terms if they are well-established. (See also 9.3.3.2 and 9.3.3.3.)

EXAMPLE SET 1 (of loan terms)

In English: bouquets gestalt therapy ombudsmen
In French: handball permafrost pizza

Occasionally a loan term and a putative translation coexist. If the loan term is more widely accepted, it should be treated as the preferred term, whereas the translation may be preferred if it has become well-established. Reciprocal references should be made between the two terms.

EXAMPLE SET 2 (of loan terms with translations)

In English: <i>coiffeurs</i> USE hairdressers hairdressers UF <i>coiffeurs</i>
abattoirs UF <i>slaughterhouses</i> <i>slaughterhouses</i> USE abattoirs
In French: baladodiffusion EP <i>podcast</i> podcast EM <i>baladodiffusion</i>

6.6.4 Transliteration

When transliterating terms from languages with different alphabets, a recognized scheme should be followed. The scheme used should be specified in the introduction to the thesaurus. Some relevant schemes can be found within Reference [27] and the standards publications listed in the Bibliography. See also 12.4, which addresses machine encoding issues associated with use of alternative scripts.

6.6.5 Neologisms, slang terms and jargon

Established terms for concepts should in general be used instead of slang or jargon terms, since these are often transient.

Useful neologisms, however, are often coined in response to emerging technologies or developments in society or politics. They may be adopted as thesaurus terms as soon as they are judged useful for the retrieval of information.

EXAMPLE SET 1 (of useful neologisms)

In English: carbon trading smart cards
In French: alicament baladodiffusion littérisme

Slang or jargon entries may be needed in circumstances such as the following:

- a) A newly-emerging concept is expressed by a term originating within a particular subculture or social group, and no widely accepted alternative exists. The slang or jargon term should then be accepted as a preferred term.

EXAMPLE SET 2 (of acceptable slang or jargon terms)

hippies web feeds

- b) Where a slang or jargon term is widely used in place of the well-established term, many users would find it a useful entry point. A non-preferred term should be provided from the slang or jargon to the preferred term.

EXAMPLE SET 3 (of slang or jargon terms given non-preferred status)

In English: psychiatrists UF <i>shrinks</i> <i>shrinks</i> USE psychiatrists
In French: <i>policier</i> EP flic <i>flic</i> EM policier

- c) Where the anticipated audience has a strong preference for slang or jargon it may be helpful to use slang/ jargon as preferred terms, entering the alternative terms as non-preferred. This could happen, for example, with an audience of young people (preferring slang).

6.6.6 Common names and trade names

A product is frequently known by a recognized trade name. Where a suitable common name also exists, this should normally be adopted as the preferred term, and the trade name should be admitted as a non-preferred term only if it is likely to serve as a user's access point. As registered trademarks enjoy legal protection, the symbol "®" should be added to the term to avoid legal problems.

EXAMPLE 1

soluble coffee
UF Nescafé®
Nescafé®
USE soluble coffee

Exceptions can occur, for example in applications where one generic product is sold under several competing trade names and the users want to distinguish between the different brands. In this case the preferred term for the generic product should be the common name, and the trade names may be used as narrower terms. See also the discussion of specificity in 8.4.

EXAMPLE 2

ballpoint pens
NT Bic pens
Biros

6.6.7 Popular names and scientific names

If a popular and a scientific name refer to the same concept, preference should be given to the form more likely to be sought by users of the thesaurus. For example, "penguins" might be chosen as the preferred term in a general thesaurus, but the scientific equivalent, "Sphenisciformes", may be preferred in a zoological thesaurus. Reciprocal references should be made in these cases.

6.6.8 Abbreviations and acronyms

Abbreviations and acronyms are often ambiguous, because they can refer to more than one concept. Therefore the full form of the name should normally function as the preferred term, with a reciprocal reference from the abbreviated form.

EXAMPLE SET 1 (of acronym/abbreviation disambiguation)

In English: United Nations Environment Programme UF <i>UNEP</i> <i>UNEP</i> USE United Nations Environment Programme direct current UF <i>DC (direct current)</i> <i>DC (direct current)</i> USE direct current national insurance UF <i>NI (national insurance)</i> <i>NI (national insurance)</i> USE national insurance	In French: nouveaux pays industriels EP <i>NPI (nouveaux pays industriels)</i> <i>NPI (nouveaux pays industriels)</i> EM nouveaux pays industriels
---	---

Exceptionally, abbreviations or acronyms may be used as preferred terms when they are widely known, unambiguous and readily understood within the field covered by the thesaurus, especially if they have become so well-established that the full form of the name is rarely used or is generally ignored. Reciprocal references should still be made between the full term and its abbreviation.

EXAMPLE SET 2 (of acronyms/abbreviations which do not need disambiguation)

In English: UNICEF UF <i>United Nations International Children's Emergency Fund</i> <i>United Nations International Children's Emergency Fund</i> USE UNICEF HIV UF <i>human immunodeficiency virus</i> <i>human immunodeficiency virus</i> USE HIV	In French: SIDA EP <i>Syndrome d'immunodéficience acquise</i> <i>Syndrome d'immunodéficience acquise</i> EM SIDA
--	---

6.6.9 Proper names

6.6.9.1 General

Proper names are often useful in a thesaurus but are sometimes excluded if there are too many candidates, especially if their forms are controlled by another authority list, or by a set of rules such as the *Anglo-American cataloguing rules*²⁾[37]. However, they should be included if the thesaurus is the only means of validating index terms. Inclusion has the additional benefit that hierarchical or associative links (see Clause 10) can be established between subject terms and relevant proper names, for example when the latter represent instances of the former.

EXAMPLES

mountains	
NT	Ben Nevis
	Mount Everest
Hockney, David (1937-)	
BT	artists
RT	contemporary art

6.6.9.2 Place names

Names of countries and geographical regions sometimes vary from language to language. Variant terms referring to the same place also occur within a single language community for reasons such as the following:

- a) a "popular" and an "official" name are both in common use;

EXAMPLE 1

Republic of Korea
South Korea

- b) the vernacular form of the name differs from the usual form of the name in the language of the thesaurus;

EXAMPLE 2

Leghorn	(A thesaurus in English may include this translation of the Italian city locally known as "Livorno")
Livorno	(The same thesaurus may also include the vernacular name)

- c) in a country with more than one official language, two forms may commonly be used;

EXAMPLE 3

Gand	(In Belgium, this French name for the city of Ghent coexists with the name in Dutch)
Gent	(In Belgium, this Dutch name for the city of Ghent coexists with the name in French)

2) While this set of rules is widely used at the time of preparation of this part of ISO 25964, there are plans to replace it with a new standard known as RDA: Resource Description and Access.

The name that is most familiar to the users of the thesaurus should be designated as the preferred term. Preference should be given to the official rather than the popular name. The short form of the official name should be preferred. Standard authorities, such as ISO 3166, should be consulted for the official forms. Reciprocal references should be made between the preferred and non-preferred forms.

EXAMPLE 4

Ireland	
UF	<i>Irish Republic</i>
	<i>Eire</i>
<i>Eire</i>	
USE	Ireland
<i>Irish Republic</i>	
USE	Ireland

6.6.9.3 Names of institutions and persons

Variant names are common and can create problems if they are not controlled. When included in a thesaurus, the form of the names should be selected in accordance with a recognized code of cataloguing practice, such as the *Anglo-American cataloguing rules*³⁾[37]. Care should be taken to ensure the name used as the preferred term is correct, current and sufficiently complete to avoid confusion with other persons or institutions. Alternative forms of the name that are in common use should be established as non-preferred terms.

EXAMPLES

Armstrong, Louis (1901-1971)	UF	Satchmo (1901-1971)
Berners-Lee, Tim		
Otlet, Paul (1868-1944)		
Centre national de la danse (France)	UF	CND (Centre national de la danse)
Intergovernmental Committee for Physical Education and Sport		
Association for Professional Broadcasting Education (U.S.)	UF	Broadcast Education Association (U.S.)

7 Complex concepts

7.1 General

Concepts range from the very simple to the very complex. At the simple end are notions such as "silver" or "people". Greater complexity is found in "Georgian silver teapots" or "human rights campaigners". The latter expression, for example, combines at least three distinct concepts into one more complex concept.

Complex concepts are very often conveyed by compound terms, i.e. terms that can be morphologically split into two or more components. Some of these are multi-word terms, and some consist of just one word.

"Biodegradability", for example, is a single-word term combining the idea of being easy to break down with the notion that living organisms will be the cause of the breakdown. In theory one could synthesize the concept with a combination of "degradability + living organisms".

3) While this set of rules is widely used at the time of preparation of this part of ISO 25964, there are plans to replace it with a new standard known as RDA: Resource Description and Access.

Splitting "biodegradability" into two concepts might seem cumbersome or unsatisfactory, but when the complex concept is usually conveyed by a multi-word term, splitting into simpler concepts becomes a more easily understandable option. For example, "human rights campaigners" can easily be split into "human rights + campaigners", or "camping holidays" could be adequately conveyed using "camping + holidays". Some concepts present many options for synthesis. For example, "human resource management" could be conveyed by "human resources + management" or "people + resource management" or even "people + resources + management".

The availability of so many choices presents the thesaurus editor with a difficult and subjective decision: whether to admit the complex concept or whether to rely on simpler concepts for users to apply in combination. Where the thesaurus allows several options, an important consideration is to guide all users to use the same combination for a given concept. Consistency becomes even harder to achieve if the concept is split into more than two components, as in the example of "people + resources + management".

Another consideration is that the introduction of complex concepts, whether represented by single-word or multi-word terms, tends to increase specificity. Greater specificity helps users to achieve greater discrimination. For example, if "biodegradability" is admitted as a narrower term of "degradability", searchers have an easy way to find relevant documents without unwanted items on degradability by chemical means, the action of ultraviolet light, etc. However, there is a penalty in that the thesaurus becomes larger and sometimes separates like items, making it harder for users to identify the right term.

Most of the examples in Clause 7 are based on multi-word terms. However, similar considerations apply to many complex concepts conveyed by single-word terms. Furthermore, in languages such as German, which uses very many compound words, the option of splitting single-word terms will occur much more commonly than in English and the same guidance applies.

7.2 The nature of compound terms

In the English language, compound terms very often consist of more than one word.

EXAMPLE SET 1

road safety thesaurus management software birds of prey

The parts of most such terms can be distinguished as follows.

- a) The focus or head, i.e. the noun component that identifies the general class of concepts to which the term as a whole refers.

EXAMPLE SET 2 (of term focus)

- | |
|--|
| 1) the noun component "indexes" in the term "printed indexes" |
| 2) the noun "hospitals" in the prepositional phrase "hospitals for children" |

- b) The difference or modifier, i.e. one or more further components that serve to narrow the scope of the focus and thereby specify one of its subclasses.

EXAMPLE SET 3 (of term difference)

- | |
|--|
| 1) the adjective "printed" in the term "printed indexes" |
| 2) the preposition-plus-noun combination "for children" in the term "hospitals for children" |

A similar analysis can sometimes be applied to single-word compound terms.

EXAMPLE SET 4 (of single-word compound terms)

Term	Focus	Difference
agroindustry	industry	agro (derived from "agriculture")
biosecurity	security	bio (prefix associated with biology)

7.3 Deciding whether or not to admit a complex concept

7.3.1 The options in outline

Taking a straightforward example, a concept such as "road safety" could be represented using a combination of the simpler concepts "roads" and "safety". As the complexity increases, more choices have to be considered. Thus, thesaurus management software could be represented by "thesauri" + "management" + "software" or by "thesaurus management" + "software".

The following five main options should be considered for handling a proposed complex concept.

- a) Admit the concept, representing it with a single preferred term.

EXAMPLE SET 1

camping holidays road safety thesaurus management software
--

- b) Admit the concept, representing it as a combination of two or more terms. In this case, a non-preferred term for the whole concept should be entered, pointing to the appropriate combination of simpler preferred terms. This option is often referred to as "splitting" the concept.

EXAMPLE SET 2

camping holidays	USE	camping + holidays
thesaurus management software	USE	thesaurus management + software

- c) Admit the concept as a candidate, for review after a trial period.
- d) Reject the complex concept, as too infrequently required, irrelevant or inappropriate.
- e) Reject the concept for direct inclusion, but retain a broader concept and admit a non-preferred term pointing to the broader preferred term.

EXAMPLE SET 3

road safety	USE	safety
thesaurus management software	USE	software

7.3.2 Factors to consider

The decision on admission is often difficult and subjective. It may in part be guided by analysis of any transaction log files that are available (see 13.1.3.3 and 13.2.2). The thesaurus developer should also consider the balance of advantages in the retrieval situations that are likely to occur. Factors such as the following (which are not listed in order of priority) should be borne in mind.

- a) Frequency of use. Some provision should be made for a candidate complex concept if the concept is frequently sought or if it is of growing importance to the user community. This is all the more straightforward if the term representing it is widely used and understood by the audience. Option a) of 7.3.1 should usually be adopted but options b) and c) may also be considered.
- b) Degree of discrimination needed. If there is a large anticipated volume of relevant documents to be indexed, it may be important to discriminate between the complex concept and similar concepts. Where "road safety" occurs, for example, the same document collection may also deal with rail safety and/or transport safety more generally, as well as fire safety, safety in the home, etc. Many users might want to discriminate between these different aspects of safety. Admitting all these terms gives an increased level of specificity, allowing greater precision in both indexing and searching.
- c) Multiplicity of search options, leading to incomplete retrieval. The availability of a number of closely-related preferred terms complicates the choice of options. For example, if the thesaurus has all of the terms "safety", "transport safety", "passenger safety", "pedestrian safety", "vehicle safety" as well as "road safety", the searcher might need to try all these options (and also combinations of "safety" with terms such as "roads", "vehicles", "cars", "road traffic", etc.) to be sure of finding all the relevant material. If the total number of documents on safety is small, searching might be easier and more effective if none of the complex concepts is accepted as a preferred term and a post-coordinated search expression such as "roads + safety" is the only retrieval option.
- d) Compound terms in which one focus has more than one difference. In the preceding examples, relatively simple terms were used in which the focus "safety" was qualified by only one difference, for example "transport". Greater complexity occurs with terms such as "vehicle fire safety", for which the focus is qualified simultaneously by two differences. Terms like these multiply still further the search options and impede straightforward retrieval, and so they should be avoided. For example, the combination of "vehicle safety" + "fire safety" adequately represents the needed concept and causes no confusion.
- e) Expectations of the anticipated users. A term such as "pre-coordinate indexes" could be acceptable and useful in a thesaurus to serve information retrieval specialists, but could be confusing if the thesaurus is intended for a more general audience. In the latter case the best solution might be not to split the term into "pre-coordination" + "indexes", but to choose option e) of 7.3.1, making it a lead-in term to the broader term "indexes" (see 8.4).
- f) Avoiding ambiguity. Sometimes inclusion of a complex concept is needed to avoid ambiguity if the component concepts could be combined in different ways, having different meanings. For example, a combination of "libraries" and "science" could be used to represent "library science" or "science libraries". A preferred term for one or both of the complex concepts may, therefore, be admitted to avoid the retrieval of unwanted items.
- g) Multilingual complications (see Clause 9). In the case of multilingual thesauri, constraints or particular needs found in any one of the languages may influence inclusion or exclusion of a complex concept across all the languages.
- h) Anticipated search environment. If the search software offers sophisticated capabilities such as clustering of search results, customization to reflect the interests of particular users or user communities, etc., the level of specificity needed in the thesaurus can be affected. This factor should be balanced against the probability that the software and other aspects of the environment might change in future.

7.3.3 Circumstances that favour splitting a complex concept

The following circumstances favour splitting.

- a) Splitting may be considered when the concept is quite specific and falls outside the core scope of the thesaurus. Inclusion of a large number of peripheral terms increases the bulk and complexity of the vocabulary without giving much retrieval benefit.
- b) If very few documents are likely to be indexed with the proposed term, inclusion as a preferred term might not be worthwhile, and splitting is an option to consider.
- c) The complex concept should be split if the focus is qualified by more than one difference. For example, "underwater cine cameras" should be split into "underwater cameras" and "cine cameras".
- d) The complex concept should usually be split if the focus represents a property, part or component of the difference.

EXAMPLES

aircraft engines hospital floors instrument reliability soil acidity

However, exceptions can occur if the concept is well-defined and distinctive, for example "lamp shades" or "body temperature". References to body temperature are likely to differ significantly from references to the temperature of other inanimate objects or substances that might be discussed in the same document collection.

7.3.4 Circumstances that rule against splitting a complex concept

A complex concept should not be split if the following conditions apply.

- a) A term for the concept has become so familiar in common use, or in the field covered by the thesaurus, that its expression as separate elements would hinder comprehension.

EXAMPLE SET 1

data processing gross domestic product

- b) Splitting the concept into its parts would lead to a loss of meaning, or ambiguity.

EXAMPLE SET 2

In English: plant food ("food" + "plants" could represent either "food" for "plants" or "plants" as "food".)
In French: marque de voiture ("marque" + "voiture" could represent either a "voiture" used as a "marque" or a "marque" of a "voiture".) société de construction ("société" + "construction" could represent either a "société" for "construction" or the "construction" of a "société".)

- c) The concept is represented by an established term that is a proper name, or incorporates a proper name.

EXAMPLE SET 3

Boolean logic United Nations

- d) The difference in an established term has lost its original meaning.

EXAMPLE SET 4

In English: lawn tennis deck chairs trade winds
In French: pistolet à température maison mère police d'assurance

- e) The concept is represented by a term containing a difference suggesting a resemblance, as a metaphor, to an unrelated thing or event.

EXAMPLE SET 5

tree structures wing nuts

- f) The parts of the concept when separated do not convey the overall concept. Thus, it would be misleading to index the corresponding documents with terms representing the part concepts.

EXAMPLE SET 6

In English: fire escapes swimming pools	(a discussion of fire escapes might say nothing about fires or escaping) (a discussion of swimming pools might have little information on swimming)
In French: établissement de santé classe de neige	(a discussion of "établissement de santé" might have little information on "santé") (a discussion of "classe de neige" might have little information on "neige")

- g) The concept is represented by a term in which the focus has another meaning in the absence of the difference.