INTERNATIONAL STANDARD

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)

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

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

hounds 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
Descriptive elements		**SN**	Scope note
		DEF	Definition
		HN	History note
Codes		**SC**	Subject category; a code or notation applied to a group of concepts relevant to a particular subject
		CC	Concept code or notation
Relationships	→	**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] *Note d'emploi*	H Erläuterung (Hinweis)	SN Forklarende note	Huomautus[c]	NO Note	SN Anmärkning	NA Nota de aplicación *Nota de alcance*	J 解 jie	лп Лексическое примечание
USE Use	EM Employer	BS Benutze	USE Brug	KÄ käytä	BRUK Bruk	USE Använd	USE Use	Y 用 yong	см Смотри
UF Use for *Used for*[a]	EP Employé pour *Employer pour*	BF Benutzt für *Benutze für*	UF Brugt for	KT korvattu termi	BF Brukt for	UF Används istället för	UP Usado por *Use por*	D 代 dai	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 shu	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 fen	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 can	а Ассоциация
BTG Broader term (generic)	TGG Terme générique (générique)	OA Oberbegriff (Abstraktionsrelation)	BTG Overordnet term (generisk)						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 (generisk)						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 **UF** *automobiles*	*automobiles* **USE** cars
animals **UF** *fauna*	*fauna* **USE** animals
World Health Organization **UF** *WHO*	*WHO* **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.

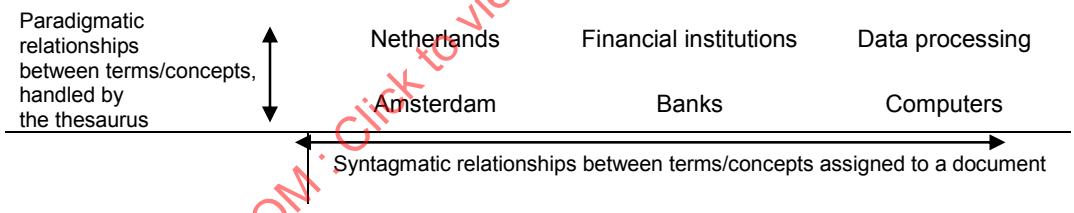


Figure 1 — Paradigmatic and syntagmatic relationships

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:

a) the equivalence relationship as it applies in both monolingual and multilingual situations (see Clauses 8 and 9 respectively);

b) the hierarchical relationship (see 10.2);

c) 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	*rather than*	the arts
state (political entity)	*rather than*	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

Salvador (country)	**USE**	El Salvador
The Needles	**USE**	Needles (Isle of Wight)
Who, The	**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	*Not*	β-rays
databases	*Not*	data-bases
nonfiction	*Not*	non-fiction
research and development	*Not*	research & development
In French:		
rayon bêta	*au lieu de*	rayon β
recherche et développement	*au lieu de*	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	**USE**	nonfiction
résumés	**USE**	resumes
In French:		
contre-marche	**EM**	contremarche
contre-plaqué	**EM**	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:		
Roumania	**USE**	Romania
Rumania	**USE**	Romania
In French:		
Lithuanie	**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**	*color*
defence	**UF**	*defense*
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 *abatoirs*
abbatoirs
abbattoirs

In French:

Mitterrand François
EP *Mitterand François*

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: *coiffeurs* **USE** hairdressers hairdressers **UF** *coiffeurs*
abattoirs **UF** *slaughterhouses* *slaughterhouses* **USE** abattoirs
In French: baladodiffusion **EP** *podcast* podcast **EM** *baladodiffusion*

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** *shrinks* *shrinks* **USE** psychiatrists
In French: *policier* **EP** flic *flic* **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** *UNEP* *UNEP* **USE** United Nations Environment Programme direct current **UF** *DC (direct current)* *DC (direct current)* **USE** direct current national insurance **UF** *NI (national insurance)* *NI (national insurance)* **USE** national insurance
In French: nouveaux pays industriels **EP** *NPI (nouveaux pays industriels)* *NPI (nouveaux pays industriels)* **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** *United Nations International Children's Emergency Fund* *United Nations International Children's Emergency Fund* **USE** UNICEF HIV **UF** *human immunodeficiency virus* *human immunodeficiency virus* **USE** HIV
In French: SIDA **EP** *Syndrome d'immunodéficience acquise* *Syndrome d'immunodéficience acquise* **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 *Irish Republic*
Eire

Eire
USE Ireland

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

EXAMPLE SET 7

artificial flowers	
chocolate eggs	(It would be misleading to use the term "eggs" to index items dealing with chocolate eggs, since the chocolate ones are a type of confectionery rather than a type of egg.)

7.4 How to split a complex concept

In simple cases, for example that of road safety, the concept may be adequately expressed using the words in the multi-word term, i.e. "roads" + "safety". The justification for doing this in a particular thesaurus might be that any document dealing with road safety is likely to contain information relevant to both of the constituent terms and that somebody looking for information about roads is likely to judge an article on road safety as relevant.

When a concept is split, the constituent concepts may be represented by preferred terms that differ from the words of the multi-word term. Often the constituent words need to be modified to convey the correct concept. For example, "rail safety" should not be split into "rails" + "safety", but might be expressed as "railways" + "safety". For "human resources" the combination of "people" + "resources" is probably more acceptable to users than "humans" + "resources".

For complex concepts that do not justify a single preferred term but are likely to be sought by users, thesaurus entries of the following form should be provided (see 8.5 for more details).

EXAMPLE

rail safety	
USE	railways
+	safety

7.5 Retention of constituent concepts

7.5.1 General

In the event that a compound term is accepted as a preferred term, consideration should be given to inclusion of preferred terms representing each of its constituent concepts if they are not already present in the thesaurus. For example, if "road safety" is accepted, the thesaurus should usually also contain the preferred terms "roads" and "safety". The latter term would be established as a broader term of "road safety", and "roads" as a related term (see 10.3).

7.5.2 Parts and components

Confusion can arise when the difference of the compound term is a machine or other complex assembly of which the focus refers to a component. For example, if "aircraft engines" is split into "aircraft" and "engines", a search for the term "aircraft" on its own will return records of all aircraft components as well as complete aircraft.

This might be acceptable for some collections, but in others will result in an overloading of the term "aircraft" and provide no means of isolating the documents dealing with whole aircraft. One solution is to create a preferred term such as "aircraft components". This term may be used in combination with "engines", or "instrumentation" or "under-carriages" or any other type of component.

7.6 Consistency in the treatment of complex concepts

In general, consistency of practice is a useful aim as it encourages consistent application by users. However, absolute consistency in the admission of complex concepts is difficult to achieve and is not always necessary. As the specificity needed in core areas of the thesaurus is usually greater than that for peripheral areas, some inconsistencies are inevitable. This type of inconsistency does not have a negative effect on retrieval performance as long as there are enough clear and helpful entries in the thesaurus, and so it may be disregarded.

In some circumstances, however, consistent rules are helpful for thesaurus editors. For example, when a thesaurus is shared among several organizations, using it for different document collections and user groups, judgements about the level of specificity are harder for one person to make. Similarly, if thesaurus compilation is a networked project, in which suggestions are invited from hundreds of contributors around the world, some guidance is necessary.

To achieve consistency, a set of criteria appropriate to the subject field should be established. For example, the *Art & Architecture Thesaurus*[26] has a set of rules for when to split compound terms, with an emphasis on the needs in cataloguing objects in museums and galleries. One such rule is to split a compound if the difference is a style or period term, thus the concept "baroque gilding" would be represented by the combination "baroque" + "gilding".

7.7 Order of words in multi-word terms

Preferred terms that are either adjectival or prepositional noun phrases should be entered in the thesaurus in natural language order, not as inverted terms. However, the inverted form of a prepositional phrase may be entered as a non-preferred term. This is especially useful if the thesaurus is used in printed form; it might not be necessary for an electronic thesaurus that can be searched for any word embedded in a term.

EXAMPLES

matter, states of	**USE**	states of matter
prey, birds of	**USE**	birds of prey

For adjectival phrases (for example "brown bread" or "double beds") an inverted entry might not be necessary because the focal noun is usually present as a broader term (see 10.2). The entry for the broader term (for example "bread" or "beds") provides an entry point from which all its narrower terms can be found.

8 The equivalence relationship, in a monolingual context

8.1 General

The equivalence relationship is the relationship between a preferred term and its corresponding non-preferred term(s) in the same natural language. As shown in the data model in Clause 15, this relationship applies between terms rather than between concepts.

NOTE The equivalence relationship between terms representing the same concept in different natural languages, often referred to as cross-language equivalence, is discussed in Clause 9.

The relationship is reciprocal, and reciprocity between preferred and non-preferred terms should be expressed by the following conventions:

USE, written as a prefix to the preferred term;

UF (use for *or* used for), written as a prefix to the non-preferred term.

EXAMPLE

In English: greenhouses **UF** *glasshouses* *glasshouses* **USE** greenhouses
In French: biodiversité **EP** *diversité biologique* *diversité biologique* **EM** biodiversité

Equivalence is established in four general situations:

a) the terms are synonyms (see 8.2);

b) the terms are quasi-synonyms (see 8.3)

c) the term is regarded as unnecessarily specific and it is represented by another term with broader scope (see 8.4);

d) the term is regarded as unnecessarily specific and it is represented by a combination of two or more terms (known as "compound equivalence"). (See 8.5).

8.2 Synonyms

Various types of synonym are encountered in practice. The following list is not exhaustive, but it indicates some of the more common classes of synonyms.

a) Terms of different linguistic origin.

EXAMPLE SET 1

In English: freedom; liberty sweat; perspiration
In French: marketing; mercatique lobby; groupe de pression

b) Popular names and scientific names.

EXAMPLE SET 2

In English: baking soda; sodium bicarbonate rock roses; Cistus
In French: aspirine; acide acetylsalicylique ail; Allium sativum

c) Common nouns and trade names.

EXAMPLE 3

vacuum flasks; Thermos® flasks; Thermos® bottles

d) Variant names for emergent concepts.

EXAMPLE SET 4

hovercraft; air cushion vehicles laptop computers; notebook computers

e) Current or favoured terms versus outdated or deprecated terms.

EXAMPLE SET 5

developing countries; underdeveloped countries radio; wireless

f) Variant spellings, including stem variants, inverted word order and irregular plurals. Among the variant spellings, it might be useful to include some misspellings (see 6.6.2).

EXAMPLE SET 6

In English: geese; goose groundwater; ground-water; ground water paediatrics; pediatrics radiation, ionizing; ionizing radiation; ionising radiation Romania; Rumania; Roumania
In French: clef; clé Lithuanie; Lituanie nénuphar, nénufar bioénergie, bio-énergie oeil, yeux

g) Terms originating from different cultures sharing a common language.

EXAMPLE SET 7

In English: flats; apartments lifts; elevators
In French: mitaine (fr-CA); moufle (fr-FR) journal de classe (fr-BE) ; agenda scolaire (fr-FR)

h) Abbreviations or acronyms and full names.

EXAMPLE SET 8

FAO; Food and Agriculture Organization pvc; polyvinyl chloride

i) Common nouns and slang or jargon terms.

EXAMPLE SET 9

In English: psychiatrists; shrinks soluble coffee; instant coffee
In French: policier; flic café soluble; café instantané

In these and similar cases, preferred terms should be selected to serve the needs of the majority of users, bearing in mind the recommendations in 6.4 and 6.6. For the sake of predictability, these criteria should be applied consistently throughout the thesaurus. If, for example, it is decided that popular names rather than scientific names should serve as preferred terms, this decision should be applied consistently except when an appropriate popular name does not exist. Similarly, consistent spelling conventions should be adopted. The editorial conventions and criteria should be noted in an introduction to the thesaurus (see 13.4).

As an option, it is possible to designate which type of equivalence prevails between a pair of terms, by using tags other than USE/UF. The following tags are sometimes used (in English).

SP SPelling variant

MS MisSpelling (see also 6.6.2)

AB ABbreviation

FT Full form of the Term

Additional tags may be devised to extend this principle to the cases described in 8.3 and 8.4. However, customization should only be undertaken with caution (see 10.4). All non-standard tags should be fully explained in the Introduction (see 13.4).

8.3 Quasi-synonyms

Quasi-synonyms frequently represent points on a continuum, and can even be antonyms (opposites).

EXAMPLE SET 1 (of opposites treated as quasi-synonyms)

consistency, inconsistency wetness, dryness

In the second line of Example set 1 above, the concept sought in the thesaurus is "the amount of moisture in an object or material", which could be expressed in terms either of wetness or of dryness. Arbitrarily, one of the terms is chosen as the preferred term and the other as the non-preferred term.

In other cases, the concepts are so closely related that discussion of one is likely to interest a user seeking the other.

EXAMPLE SET 2 (of related concepts treated as quasi-synonyms)

bushes, shrubs mallets, hammers

The extent to which terms are treated as quasi-synonyms depends to a large extent upon the subject field covered by the thesaurus. In a thesaurus specializing in clothing manufacture, the terms "gloves" and "mittens" might both be established as preferred terms, whereas a more general thesaurus for the textile industry as a whole might treat the two terms as quasi-synonyms. The decision should be based on anticipating the degree of discrimination required at the time of searching.

8.4 Specific terms subsumed in a broader concept

It is sometimes helpful to treat the name of a class, and also the names of its members, as an equivalence set, with the broader term functioning as the preferred term.

EXAMPLE

rock	
UF	*basalt*
	granite
	slate
	etc.
basalt	**USE** rock
granite	**USE** rock
slate	**USE** rock

This technique is sometimes used to reduce the number of preferred terms in a thesaurus. It limits the specificity of indexing and searching that can be achieved, and is beneficial when the collection to be indexed has very little information on the subject area in question. The presence of the non-preferred terms provides extra entry points to assist user access. Subsumption of narrower concepts in a broader one should not, however, be used when there is a wealth of information on the subject in question and greater specificity is needed to pinpoint concepts with accuracy.

8.5 Representation of complex concepts by a combination of terms

Where a multi-word term is deemed to be unsuitable as a preferred term, but might be sought by some users, it may be represented by a combination of two or more preferred terms [sometimes known as compound equivalence; see also 7.3.1, option b)]. An entry of the following type may be admitted in the thesaurus.

EXAMPLE SET 1 (of entries for complex concepts)

coal mining
USE coal
\+ mining

ferromagnetic films
USE ferromagnetic materials
\+ films

Reciprocals in the following form should be provided.

EXAMPLE SET 2 (of reciprocal entries for a complex concept)

coal
UF+ *coal mining*

mining
UF+ *coal mining*

The three-way relationship appears complex, but the tags **USE** and **+** are differentiated only for ease of legibility. In functional terms, the relationship between "coal" and "*coal mining*" is identical to that between "mining" and "*coal mining*", in both directions.

Hypothetically, combinations of terms could be linked with "OR" rather than "AND" (see Example set 3). However, this situation should be handled as a case of homography (see 6.2.2). The "OR" connector should be avoided, and entries that clarify the scope of the alternatives should be provided instead (see Example set 4 below).

EXAMPLE SET 3 (of practice to be avoided)

In English:

pitch
USE audio frequency
OR gradient

In French:

ton
EM nuance de couleur
OU tonalité (musique)

However, this type of usage should be avoided, and entries that clarify the scope of the alternatives should be provided instead.

EXAMPLE SET 4 (of avoidance of complex relationships involving "OR")

In English: *pitch (sound)* **USE** audio frequency *pitch (steepness)* **USE** gradient
In French: *ton (couleur)* **EM** nuance de couleur *ton (musique)* **EM** tonalité (musique)

9 Equivalence across languages

9.1 General

In a multilingual thesaurus, all languages should have equal status and, if possible, each concept should be represented in every language of the thesaurus. A collection that has been indexed using any one of the languages can then be searched equally effectively using any of the other languages.

When every concept in a multilingual thesaurus has a preferred term in each language and all languages share exactly the same structure of hierarchical and associative relationships, the structure is said to be "symmetrical". For discussion of non-symmetrical structures, see 9.2 c), 10.2.6 and 10.3.4.

It is possible also to treat different dialects or sublanguages as though they were separate languages. For example, American English, British English and Indian English may be treated as different languages, all three presented separately in one trilingual thesaurus. Many of the terms are common to two or three of the languages, but other terms are different. Similarly, the terminology preferred by scientists may be presented as a different language to that of marketing and sales personnel. If the thesaurus is treated as monolingual, one preferred term is assigned to each concept, and the alternative scientific term or dialect term appears as a non-preferred term [(see 6.6.2 and 8.2 g)]. Treating it as a multilingual thesaurus allows equal status to be given to each dialect or sublanguage.

In a multilingual thesaurus, a bi-directional cross-language equivalence relationship should be shown between preferred terms that represent the same concept in each of the languages. As an option, each concept may be given a notation and/or unique identifier (see 12.1.3).

In cases where a suitable term is not available in one language, it is often possible to modify the scope of the concept so that it can be represented by a suitable term in each of the thesaurus languages. Alternatively, an inexact or partial equivalent may be accepted in one language, or a gap in one of the languages may be filled with a loan term or a coined term (see 9.3.3).

Where it proves impossible to find suitable terms to represent every concept in each of the languages of a multilingual thesaurus, it might be necessary to treat the different language versions of the multilingual thesaurus as if they were two or more parallel monolingual thesauri and to establish mappings between the corresponding terms. The establishment of mappings will be described in ISO 25964-2.

NOTE: Approaches for managing the construction of a multilingual thesaurus are discussed in 13.3.3.

9.2 Degrees of equivalence

During preparation of a multilingual thesaurus, the following degrees of equivalence are most frequently encountered. They are not to be treated as distinct relationship types but rather as points along the spectrum of possibilities that lie between the extremes of exact equivalence and absence of equivalence.

a) **Exact equivalence.** In this ideal situation, a concept is represented in every language of a multilingual thesaurus, and it is possible to identify preferred terms that are semantically and culturally equivalent. A cross-language equivalence relationship should be established between the corresponding preferred terms.

EXAMPLE SET 1

en: sun / es: sol / fr: soleil / de: Sonne en: rail network / fr: réseau ferroviaire / de: Schienennetz / nl: spoorwegnet fr: physique / en: physics / es: física / de: Physik

b) **Inexact** or **near-equivalence.** The terms found in different languages sometimes carry a small difference in scope, perhaps due to differences of culture, connotation or appreciation (similar to the case of quasi-synonyms within one language; see 8.3.) If, however, the terms are accepted as being close enough to represent the same concept when used in the thesaurus, then a preferred term for each language should be selected, equivalence between them established, and the scope of the concept understood as including all the cultural variations.

EXAMPLE SET 2

en: information retrieval / fr: repérage d'information fr: jeu didactique / en: learning games en: legal education / fr: enseignement du droit en: farms / es: granjas

c) **Partial equivalence,** also known as **broader/narrower equivalence.** Sometimes one language has no term that accurately represents the concept for which a label is sought, but does have a term that is normally considered to represent a broader or narrower concept. If the scope of the concept represented by one term falls completely within the scope of the concept represented by the other, the terms are said to be partially equivalent. If the difference in scope is small enough, it might be acceptable to admit the terms to the thesaurus, treating the two as equivalents that both represent the same concept (similar to the case of specific terms subsumed in a broader concept, within one language; see 8.4.) A scope note should be added to the concept in one or both languages, to clarify any doubts. In Example 3, a German scope note is unnecessary, as the thesaurus concept corresponds closely to the normal meaning of "Wissenschaft"; but a scope note in English is needed, to clarify that a very broad definition of "science" should be understood when using the thesaurus.

EXAMPLE 3

de: Wissenschaft / en: science SN: Includes natural and social sciences and knowledge in general

A special case of partial equivalence can arise when one of the languages offers two or more partial equivalents that, in combination, represent the whole of the concept represented by one term in the source language. This situation is sometimes described as **"compound equivalence"**.

EXAMPLE 4

fr: sécurité / de: Sicherheit / en: safety, security

In Example 4, the concept represented by the French term "sécurité" is accurately labelled in German by "Sicherheit"; in English "security" normally represents part of its scope, while "safety" covers the other part. A term to cover both at once is hard to find in the English language. The situation can sometimes be handled by adding qualifiers and/or accepting coined terms in one or more of the languages (see examples in 9.3.2 and 9.3.4). If this type of solution is not acceptable it might become necessary to introduce non-symmetrical structures, or to treat the terms as belonging to separate monolingual thesauri, and establish one-to-many mappings between them as will be described in ISO 25964-2.

d) **Non-equivalence**. Sometimes no term in one language can be found that even partially or inexactly represents a concept that is needed in the thesaurus. This problem can usually be solved by means of a coined term or a loan term (see 9.3.3). For example, the Gaelic term "ceilidh" is often borrowed by English speakers to describe a social event originating in the Gaelic culture. Similarly, the concept of "régime pédagogique" is unique to the Canadian culture and if needed in a bilingual thesaurus might have to be represented in English by borrowing the French term itself, or by coining a term such as "pedagogic regime".

ISO 25964-2 will extend the discussion by covering equivalence mappings. These are used when identifying equivalent concepts in different vocabularies, or where non-symmetrical structures are found in a single multilingual thesaurus. When the concepts and structure provided in one language are not matched comprehensively in the other languages, mappings between equivalent concepts should be established selectively wherever possible, to enable navigation and support interoperability.

9.3 Typical problems and solutions

9.3.1 General

In the ideal situation, for each concept a preferred term can be found in each of the languages, and these terms are exact equivalents in normal discourse. Thus, it becomes possible for one-to-one relationships between preferred terms in all cases, and for all languages to share the same structure of hierarchical and associative relationships. The difficulties that occur in practice can sometimes be resolved by accepting inexact or partial equivalents as preferred terms [see 9.2 b) and c)]. But additional problems are often caused by quasi-synonyms (see 8.3) or unidentified homographs (see 6.2.2) in one or more of the languages, or by the complete absence of a term to represent a concept in one or more of the languages [see 9.2 d)]. There is rarely a unique correct solution. It is important to weigh up the merits of the alternatives in each case and consider what is best for most retrieval situations.

9.3.2 Problems caused by quasi-synonyms or homographs

If the problem is caused by homography, it can be resolved as recommended in 6.2.2, by adding qualifiers where appropriate (see Example 1, below). The same solution may also be applied even when the problem terms are much closer in meaning, to the extent that they could be treated as quasi-synonyms. In Example 2, there is enough difference between the near-synonyms to justify separation into two concepts, with qualifiers to bring out the difference in one of the languages. In Examples 3 and 4, a more acceptable solution is to treat the near-synonyms as equivalents in one language, with just one term to represent the same concept in the other language.

EXAMPLE 1

English	**French**
spring	printemps source
Solution	
spring (season) spring (water)	printemps source

EXAMPLE 2

French	English
lisibilité	legibility readability
Solution	
lisibilité (calligraphie) lisibilité (style)	legibility readability

EXAMPLE 3

French	English
énergie	energy power
Solution	
énergie	energy *UF power* *power* USE energy

EXAMPLE 4

English	French
walls	mur paroi
Solution	
walls	mur *EP paroi* *paroi* EM mur

9.3.3 Absence of an equivalent term in one or more languages

A source term without a cross-language equivalent is often more specific than other terms in the same thesaurus. If this is the case, one simple solution is to remove the concept and the source term that does not have an equivalent from the thesaurus. Where such removal is not possible, three solutions are available to preserve one-to-one cross-language equivalence relationships between preferred terms in a multilingual thesaurus.

9.3.3.1 Solution 1: Transform preferred term into a non-preferred term

In the source language, transform the preferred term that does not have an equivalent in the target language into a non-preferred term. This non-preferred term can then be linked to a more general preferred term with

equivalents in the other languages of the multilingual thesaurus. Effectively the concept has been changed to a broader one than that first considered.

EXAMPLE 1

English	**French**
teenagers	?
Solution	
adolescents *UF teenagers* *teenagers* USE adolescents	adolescent

EXAMPLE 2

English	**French**
classroom environment	?
Solution	
classroom environment USE learning environment learning environment *UF classroom environment*	 milieu d'apprentissage

9.3.3.2 Solution 2: Import the term as a loan term

Import the preferred source term into the target language as a loan term (see also 6.6.3). Native speakers of the target language should be consulted before any loan term is integrated in a multilingual thesaurus.

EXAMPLE SET 1 (of easily accepted loan terms)

German Schadenfreude	**English** Schadenfreude
Afrikaans veld	**English** veld

A loan term may also be used when its translation calls for a long definition or explanation that cannot be used effectively as a preferred term in the target language.

EXAMPLE 2 (of a loan term that needs an explanatory scope note)

English	**German**
teenagers	Teenager H: Zwischen 13 und 19 Jahren

A loan term is sometimes used in natural language, before a coined translation becomes prevalent. This commonly occurs with scientific discoveries or new technologies. If it appears that a translation is likely to become accepted, this should be adopted as the preferred term, and the loan term should be designated as a non-preferred term.

EXAMPLE 3 (of a loan term designated to be non-preferred)

German	English
Bremsstrahlung	braking radiation *UF* *Bremsstrahlung* *Bremsstrahlung* USE braking radiation

9.3.3.3 Solution 3: Create an equivalent (coined) term

Create an equivalent (coined term) in the target language. Terms should be coined only after consultation between indexers, language specialists and subject specialists. Such terms may be created in the following circumstances.

— The source language term, which represents a new concept to the users of the target language, is for some reason not acceptable as a loan term.

— The source language term has already been used as a loan term by authors writing in the target language, but the term needs to be replaced because it is deemed inappropriate or unacceptable. Until the newly-coined term has become established, the loan term should continue to appear in the thesaurus as a non-preferred term.

— In a multilingual thesaurus containing three or more languages, a concept first expressed in one of the languages is already represented by a coined term in one of the other languages. An indexer working in a third language, facing a choice between two available loan terms, might prefer instead to coin an equivalent, particularly if that concept is likely to occur in the third language.

EXAMPLE 1

German	English	French
Schlüsselkind	latchkey children	enfant à clé **NE** Enfant dont les parents travaillent pendant la journée et qui est muni d'une clé pour pouvoir rentrer chez lui en sortant de l'école.

Coined terms can be created in the following ways (not in order of preference).

a) By literal translation of the source language term or its semantic components.

EXAMPLE 2

French	English
cuisine minceur	lean cuisine

b) By construction of a term or phrase that expresses the general meaning of the source language term.

EXAMPLE 3

German	English	French
Bremsstrahlung	braking radiation	rayonnement de freinage

c) By invention of a neologism, which should be as concise as possible to foster acceptance (these inventions sometimes approximate to literal translations).

EXAMPLE 4

English	French
gender mainstreaming	intégration de la dimension de genre

9.3.4 Combined problems

Multiple problems often arise in one situation, and a solution might be quite complex.

EXAMPLE

Spanish	English
aislamiento	insulation isolation
Solution	
aislamiento acústico	acoustic insulation *UF insulation (acoustic)*
aislamiento eléctrico	electrical insulation *UF insulation (electrical)*
aislamiento físico	isolation
aislamiento térmico	thermal insulation *UF insulation (thermal)*

In Spanish, "aislamiento" can be considered a homograph, since it has the two meanings indicated. But separate terms do not exist for the two concepts, and qualifiers to make the difference clear are very hard to find. In English too, it can be argued that the term "insulation" is homographic, since the different types of insulation are quite different from each other. The solution illustrated therefore removes the broad concept of insulation altogether, replacing it with three narrower concepts to cover the main types of insulation needed in this thesaurus. At the same time a term "aislamiento físico" is coined in Spanish, corresponding to the English term "isolation".

9.4 Representation of cross-language equivalence between preferred terms

The alpha-2 language codes in ISO 639-1, which always use lower-case characters, should be applied to show equivalence between corresponding preferred terms in different languages (see also 15.2.10), as illustrated in the example below. If an appropriate code is not found in ISO 639-1, an alpha-3 code from ISO 639-2 may be used. These codes may be extended where necessary with the additional codes described in RFC 4646[45] and listed in the IANA subtag registry[35].

EXAMPLE

English version	Spanish version	French version
economic aid **es** ayuda económica **fr** aide économique **UF** *economic assistance*	ayuda económica **en** economic aid **fr** aide économique	aide économique **en** economic aid **es** ayuda económica

The example above shows the preferred terms and equivalence relationships for one concept in each language version of a multilingual thesaurus (English, Spanish and French). Cross-language equivalents are shown before intra-language equivalents. See 12.3 for additional guidance on the display of relationships in a multilingual thesaurus.

9.5 Cross-language equivalence between non-preferred terms

The number of synonyms or quasi-synonyms for a term usually varies from one language to another, and it is not necessary in a multilingual thesaurus to show cross-language equivalents for non-preferred terms. In the example in 9.4, "economic assistance" has been entered as a synonym for "economic aid". However, no synonyms have been entered for the Spanish and French preferred terms "ayuda económica" and "aide économique" respectively.

Showing cross-language equivalents for non-preferred terms in the different languages of a multilingual thesaurus can be useful when the non-preferred terms are quasi-synonyms conveying a specific aspect of the scope of the concept. Non-preferred terms are sometimes established for concepts at the boundaries of the subject scope of the thesaurus because the quantity of information to be described does not warrant the inclusion of a narrower concept with its own preferred term. As and when the scope of the thesaurus expands, the status of these terms can change so that they become preferred terms. When this happens, it is convenient if equivalents in all languages have already been identified.

For example, the term "narcissus" could be a preferred term in both the English and the French versions of a multilingual thesaurus, with "daffodils" and "jonquille" as non-preferred terms for the same concept, in the respective languages. It is possible to establish direct correspondence between "daffodils" and "jonquille", although this is not usually necessary.

10 Relationships between concepts

10.1 Introduction

Apart from the equivalence relationship there are various other types of relationships that can be established in a thesaurus. While equivalence relationships in a monolingual thesaurus apply only between (preferred and non-preferred) *terms*, the relationships introduced in this chapter are established between *concepts* (see also the data model in Clause 15). Such relationships should be included in a thesaurus only if they are paradigmatic, i.e. they have validity in a wide variety of contexts.

NOTE Such relationships are hard to display, since concepts exist only in the mind, and therefore the conventional way to display them is to show each such relationship as applying between the preferred terms that represent the concepts in question. Likewise the tags BT, NT and RT, while they stand for broader term, narrower term and related term respectively, can equally be taken as representing the notions of broader concept, narrower concept and related concept.

10.2 The hierarchical relationship

10.2.1 General

The hierarchical relationship should be established between a pair of concepts when the scope of one of them falls completely within the scope of the other. It should be based on degrees or levels of superordination and subordination, where the superordinate concept represents a class or whole, and subordinate concepts refer to its members or parts.

The following tags should be used, reciprocally:

BT (i.e. broader term), written as a prefix to the superordinate term;

NT (i.e. narrower term), written as a prefix to the subordinate term.

EXAMPLE 1

animals
NT mammals

mammals
BT animals

The hierarchical relationship may be one of three types, corresponding to three logically different situations as follows:

a) the generic relationship;

b) the hierarchical whole-part relationship;

c) the instance relationship.

Each of these leads to hierarchies that are amenable to a logical test through reference to the types of concept involved, for example those listed in 5.1.2. Every subordinate concept should belong to the same inherent category as its superordinate concept, i.e. both the broader and narrower term should represent a thing, or an action, or a property, etc. (See additional discussion of facet analysis in Clause 11.)

EXAMPLE 2

a) "metals" (a class of materials) and "casting" (an action) represent different types of concept and, therefore, cannot be related hierarchically;

b) "metals" and "gold" both represent materials and could, therefore, be hierarchically related.

It is possible to distinguish explicitly among the three types of hierarchical relationship using the conventions described in 10.2.2, 10.2.3 and 10.2.4. The extra work and complexity required in doing so should be balanced against the benefits for the application anticipated.

The main function of hierarchical relationships is to help both indexers and searchers choose the appropriate level of specificity. A search can be broadened or narrowed by moving up or down, respectively, in the hierarchy. A common technique for improving search recall is called "search explosion", in which the search is extended to include all of the narrower terms of the preferred term first selected [see also 16.3.3 e), 16.3.3 f) and 16.3.4 c)]. Careful adherence to the guidelines in 10.2.2 to 10.2.5 ensures that exploded searches retrieve only items that fall within the scope of the superordinate concept.

NOTE Search explosion, unlike search expansion, does not extend to associatively related terms (see 10.3).

10.2.2 The generic relationship

10.2.2.1 The generic relationship is the link between a class or category and its members or species. In addition to the test for validity described in 10.2.1, this relationship is also amenable to a logical "all-and-some test", as shown in Figure 2.

Figure 2 — Relationship that satisfies the "all-and-some test"

Figure 2 indicates that some members of the class "birds" are known as "parrots", and all "parrots", by definition and irrespective of context, are regarded as "birds". This test usually ensures that a term such as "parrots" is not subordinated to a class such as "pets", as not all parrots are pets. Figure 3 shows the relationship between the latter pair of terms.

Figure 3 — Relationship that does not satisfy the "all-and-some test"

In Figure 3, some members of the class "pets" are "parrots", and only some parrots are regarded as pets. These terms should, therefore, not be given a BT/NT linkage.

NOTE When indexing a work on "parrots as pets", preferred terms representing both of these concepts should be assigned.

10.2.2.2 This argument might not apply in the context of a specialist thesaurus devoted to domestic animals, in which the only parrots in the frame of reference are pets. In such a case, "parrots" could be subordinated to "pets" in the same hierarchy. Such approximations should be applied with caution, however, especially in networked environments in which the records from one system may be mixed with those of another. When interoperability is required, the relationships established should be universally acceptable.

10.2.2.3 The tags BT/NT are normally adequate to identify the generic relationship, but optionally, the following tags may be used:

BTG: Broader term (generic)

NTG: Narrower term (generic)

EXAMPLE

rats
BTG rodents

rodents
NTG rats

10.2.3 The hierarchical whole-part relationship

10.2.3.1 The hierarchical whole-part relationship covers a limited range of situations in which a part of an entity or system belongs uniquely to a particular possessing whole. This applies to four main classes of terms.

a) Systems and organs of the body.

EXAMPLE 1

cardiovascular system
  blood vessels
    arteries
    veins

b) Geographical locations.

EXAMPLE 2

Canada
  Ontario
    Ottawa
    Toronto

c) Disciplines or fields of discourse.

EXAMPLE 3

science
  biology
    botany
    zoology

d) Hierarchical social structures.

EXAMPLE 4

armies
  corps
    divisions (army)
      battalions
      regiments

Most other cases of the whole-part relationship are not eligible for a hierarchical linkage because the part could belong to more than one whole. For example, a BT/NT relationship should not be established between "bicycles" and "wheels" because a wheel could be part of a motor car, a wheelbarrow or one of many other artefacts. An exploded search for bicycles would retrieve much unwanted material if it were extended to all types of wheel. It is sometimes the case, however, that the parts of an artefact are unique to that artefact, at least in the field of application of the thesaurus. It might be appropriate to establish a BT/NT link between "fireplaces" and "hearths", or between "bows" and "bowstrings", for example. This would usually cause no confusion and would help with exploded searches under the broader term. However, this practice is not recommended in the case of complex machines and their components, for which a more appropriate solution is to create a broader term such as "engine components", with the various components listed as narrower terms (see 7.5.2).

10.2.3.2 The tags BT/NT are normally adequate to identify the partitive relationship, but the following tags may also be used:

BTP = broader term (partitive)

NTP = narrower term (partitive)

EXAMPLE

central nervous system **BTP** nervous system nervous system **NTP** central nervous system

10.2.4 The instance relationship

10.2.4.1 The instance relationship links a general concept, such as a class of things or events, and an individual instance of that class, which is often represented by a proper name (see also 6.6.9).

EXAMPLE

mountain regions	{ class
Alps Himalayas	{ instances

In this example, "Alps" and "Himalayas" are assigned to subordinate positions in a hierarchy. However, they are neither kinds nor parts of "mountain regions", but represent individual instances.

10.2.4.2 The tags BT/NT are normally adequate to identify the instance relationship, but the following tags may also be used:

BTI = broader term (instantial)

NTI = narrower term (instantial)

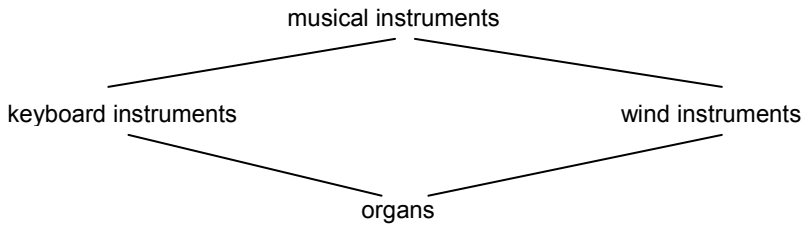
EXAMPLES

Paris **BTI** capital cities capital cities **NTI** Paris

10.2.5 Polyhierarchical relationships

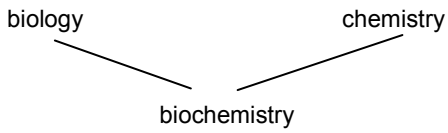
Some concepts can belong, on logical grounds, to more than one group or class at the same time. In such cases, a direct hierarchical link should be established to all of the appropriate broader concepts and the structure of the thesaurus is said to be polyhierarchical. This contrasts with a monohierarchical structure, which, for a particular concept, would allow only one of the valid BT links to be established.

EXAMPLE 1



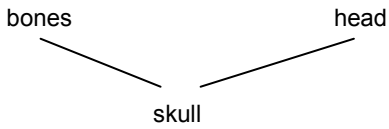
In Example 1, the term "organs" is assigned to subordinate positions on the basis of its generic relationship to two broader terms. In other cases, polyhierarchical links may be based upon whole-part relationships.

EXAMPLE 2



In some cases, polyhierarchical links can be based upon logically different relationships.

EXAMPLE 3



In Example 3, the link between "bones" and "skull" is based upon the generic relationship (the skull is a kind of bone), whereas the link between "head" and "skull" is based upon the hierarchical whole-part relationship (the skull is a part of the head).

Where one concept has more than one broader concept, all the applicable relationships should be displayed. As an option, the difference in relationship type may be explicitly shown by using the appropriate tags.

EXAMPLE 4

skull **BT** bones head	*or*	skull **BTG** bones **BTP** head

10.2.6 Non-symmetrical structures in multilingual and multicultural thesauri

Normally the same structure for a given hierarchy of concepts is shared by all the language versions of a multilingual thesaurus. However, difficulties in establishing an appropriate hierarchical structure can arise when two or more culturally different communities share one thesaurus, particularly if the concepts and terms familiar to one community do not have a one-to-one correspondence with equivalent concepts and terms used by the others [see also 9.2 c)]. To accommodate cultural and linguistic differences, and to ensure that all the language versions of one multilingual thesaurus have equal status, non-symmetrical structures may sometimes be introduced. See 13.3.3 for a discussion of approaches to building a multilingual thesaurus.

NOTE If non-symmetrical structures are admitted, the data model as described in Clause 15 has to be modified.

10.3 The associative relationship

10.3.1 General

The associative relationship covers associations between pairs of concepts that are not related hierarchically, but are semantically or conceptually associated to such an extent that the link between them needs to be made explicit in the thesaurus, on the grounds that it may suggest additional or alternative terms for use in indexing or retrieval. The relationship is indicated by the tag "**RT**" (related term) and it should be applied reciprocally.

EXAMPLE

birds
RT ornithology

ornithology
RT birds

A general guideline for associative relationships is that whenever the preferred term for one concept is used, the other should always be implied within the common frames of reference shared by the users of the thesaurus. Moreover, one of the terms is often a necessary component in any explanation or definition of the other; the term "birds", for example, forms a necessary part of the explanation of "ornithology".

It is particularly important to establish an associative relationship between concepts that overlap in scope.

10.3.2 Terms and concepts with overlapping meanings

10.3.2.1 In ordinary discourse a pair of terms can often be used interchangeably in some contexts but not in others. For example, the terms "ships" and "boats" may be said to represent different concepts because while they have much in common, neither of them adequately covers the scope of the other. If two such concepts are entered in the thesaurus and represented by distinct preferred terms, it is essential to provide an associative relationship between the two. This reminds searchers that they may need to use both preferred terms to be sure of finding all the relevant material.

EXAMPLE

boats
RT ships

ships
RT boats

10.3.2.2 A pair of preferred terms with overlapping meanings may often be siblings, i.e. they share a common broader term. However, it is not necessary to interrelate all sibling terms in this way. For example, there is no need to associate terms such as "horses" and "donkeys" on the grounds that they share a common broader term, e.g. "equines", since the meanings of the terms do not overlap in this case.

10.3.3 Other cases of associative linkage

When one term is strongly implied by another, there are sufficient grounds for associating the two terms. The following groups are offered only as representative examples of typical relational situations encountered in practice.

a) A discipline or field of study and the objects or phenomena studied.

EXAMPLE 1

forestry
RT forests

forests
RT forestry

EXAMPLE 2

neurology
RT nervous system

nervous system
RT neurology

b) An operation or process and its agent or instrument.

EXAMPLE 3

temperature control
RT thermostats

thermostats
RT temperature control

EXAMPLE 4

crime investigation
RT detectives

detectives
RT crime investigation

c) An action and the product of the action.

EXAMPLE 5

weaving
RT cloth

cloth
RT weaving

EXAMPLE 6

ploughing
RT furrows

furrows
RT ploughing

d) An action and its recipient or target.

EXAMPLE 7

harvesting
RT crops

crops
RT harvesting

EXAMPLE 8

imprisonment
RT prisoners

prisoners
RT imprisonment

e) Objects or materials and their defining properties.

EXAMPLE 9

poisons
RT toxicity

toxicity
RT poisons

EXAMPLE 10

magnets
RT ferromagnetism

ferromagnetism
RT magnets

f) An artefact and its parts, if they do not qualify for the hierarchical whole-part relationship (see 10.2.3).

EXAMPLE 11

optical instruments
RT lenses

lenses
RT optical instruments

g) Concepts linked by causal dependence.

EXAMPLE 12

bereavement
RT death

death
RT bereavement

EXAMPLE 13

diseases
RT pathogens

pathogens
RT diseases

h) An object or process and its counter agent.

EXAMPLE 14

plants
RT herbicides

herbicides
RT plants

EXAMPLE 15

inflammation
RT anti-inflammatory agents

anti-inflammatory agents
RT inflammation

i) A concept and its unit of measurement.

EXAMPLE 16

electric current
RT amperes

amperes
RT electric current

j) A compound term and the noun that is its focus, if the two do not have a true hierarchical relationship.

EXAMPLE 17

model ships
RT ships

ships
RT model ships

EXAMPLE 18

fossil reptiles
RT reptiles

reptiles
RT fossil reptiles

k) An organism or substance bred or derived from another.

EXAMPLE 19

mules
RT donkeys

donkeys
RT mules

EXAMPLE 20

brass
RT copper

copper
RT brass

10.3.4 Multilingual and multicultural thesauri

In order to provide sufficient associative relationships in a multilingual thesaurus, the needs of users of all the language versions should be considered and appropriate relationships established (see also 13.3.3).

As already described in 9.1, 9.2 and 10.2.6, communities of different cultures and/or languages sharing a common thesaurus sometimes require concepts, and inter-relationships between concepts, that are unfamiliar to each other. When non-symmetrical hierarchical structures are introduced to accommodate these differences, non-symmetrical associative relationships are usually needed too.

NOTE If non-symmetrical structures are admitted, the data model as described in Clause 15 should be modified.

10.4 Customized relationships

The equivalence, hierarchical and associative relationship types described in 8, 10.2 and 10.3 are well established, widely used and generally found to be adequate in contexts whereby a human operator views a display of broader, narrower or related terms before choosing how to express a search query. Sometimes, however, there might be a need for additional or more closely specified relationship types.

In the same way that the hierarchical relationship may be optionally subdivided into generic, partitive or instantial, the equivalence and associative relationships may optionally be subdivided for a particular application. For example, abbreviations and/or acronyms may be designated separately from all the other equivalence relationships. Tags distinct from USE/UF should be assigned, for example FT/AB for Full Term and Abbreviation. Similarly, for the associative relationship, it is possible to subdivide into types such as cause and effect designated with tags such as CAUSE/EFFECT.

Before embarking on an exercise to provide more specific relationship types, the thesaurus developer should first check that the refined relationships are really necessary, and not simply an intellectually attractive construct. The purpose of the thesaurus is to serve a set of users in a given context. The benefits of the extra relationships could be outweighed by the extra complications perceived by the users. However, there may be other justifications for the extra work, if the thesaurus is wanted to interoperate with or function as an ontology. (Ontologies usually provide more specific and closely defined relationships.)

NOTE See ISO 25964-2 for recommendations on these issues.

In the event that customization is undertaken, it is important that users understand the practice, and that confusion does not arise when the customized thesaurus is handled together with a conventional one. The introduction of non-standard tags or symbols carries the risk of incompatibility in circumstances where the thesaurus in question interoperates with others. The risks can be reduced by ensuring that the introduced relationships are true subtypes of the established relationship types.

11 Facet analysis

Facet analysis is useful in generating hierarchies that conform to the rules for hierarchical relationships, as set out in 10.2, because these relationships are valid only for concepts belonging to the same general category.

The choice of facets can vary depending on the subject field, but at the highest levels it is usual to use fundamental categories such as objects, materials, agents, actions, places, times, etc. Where it is helpful to do so, these fundamental facets may be analyzed into subfacets down to the level required; for example, actions may be subdivided into intransitive *processes* such as "ripening" or "deterioration" and transitive *operations* such as "cutting" or "repairing".

NOTE An action is said to be intransitive when the "actor" does not act upon any object; whereas the actor of a transitive action does act upon an object. Thus an apple ripens by an internal process, but in cutting an apple, a knife acts upon an external object.

An example of applying facet analysis in a classified arrangement appears in Figure 4. This shows some terms from the general subject of "industries". Directly under this term are the two narrower terms "agricultural industries" and "engineering industries". One of the facets shown under "agricultural industries" is ***products***. Within this facet, only two levels of hierarchy are shown, except for the term "milk", which has been expanded to show how its narrower terms have been grouped into arrays, each preceded by a node label showing the characteristic of division. These arrays each represent a different way of subdividing the concept of milk, by fat content, by source animal, etc., and the word "by" occurs in each of these node labels. The node labels that introduce new facets, such as (*people*) and (*products*) do not contain the word "by".

It is important to note the pattern of hierarchical relationships around the node labels. Where the label shows the characteristic of division of the superordinate term, all the terms in the array following it are true narrower terms of the superordinate term. Thus, in Figure 4, "whole milk", "buffalo milk" and "sterilized milk" are all narrower terms of "milk". In contrast, where the label introduces a new facet, the terms that follow are typically not narrower terms of the preceding term. For example, "farm managers" and "cereal products", etc., are not narrower terms of "agricultural industries".

Node labels are not thesaurus terms. They are present only for the purposes of systematic display, and they do not qualify for any of the relationships described in Clauses 8 to 10. To avoid confusion, node labels should be distinguished typographically from the thesaurus terms. Generally, italics and parentheses or angle brackets are used, as shown in Figure 4. Some alternative ways of applying facet analysis are illustrated in 12.2.4 and 12.2.5.

The terms representing concepts in an array may be arranged either alphabetically or systematically. Alphabetical sequence should be used when there is no other obvious way to arrange a group of concepts. Systematic sequence should be used when it is likely to be familiar to most users, or when the arrangement helps to clarify the scope of the terms. In the example for electromagnetic radiation, the types of radiation are presented in order of increasing wavelength, as this might help some indexers in selecting the correct term(s).

EXAMPLE

electromagnetic radiation
  <by wavelength>
  ultraviolet radiation
  visible radiation
  infrared radiation
  microwave radiation
  radio waves

NOTE This example uses slightly different but equally acceptable conventions in the node label: angle brackets instead of round brackets, and the parent term is not spelt out at the start of the node label. The presence of the word "by", however, indicates that the node label specifies the characteristic of division by which the types of radiation are differentiated.

In some thesauri, facet names are included as preferred terms and treated as top terms, under which complete hierarchies may be shown (see 12.2.4). Alternatively there may be no explicit display of complete facets and the facet names appear only in node labels, as in Figure 4, or as the names of concept groups (see 15.2.18). See also 15.2.19.

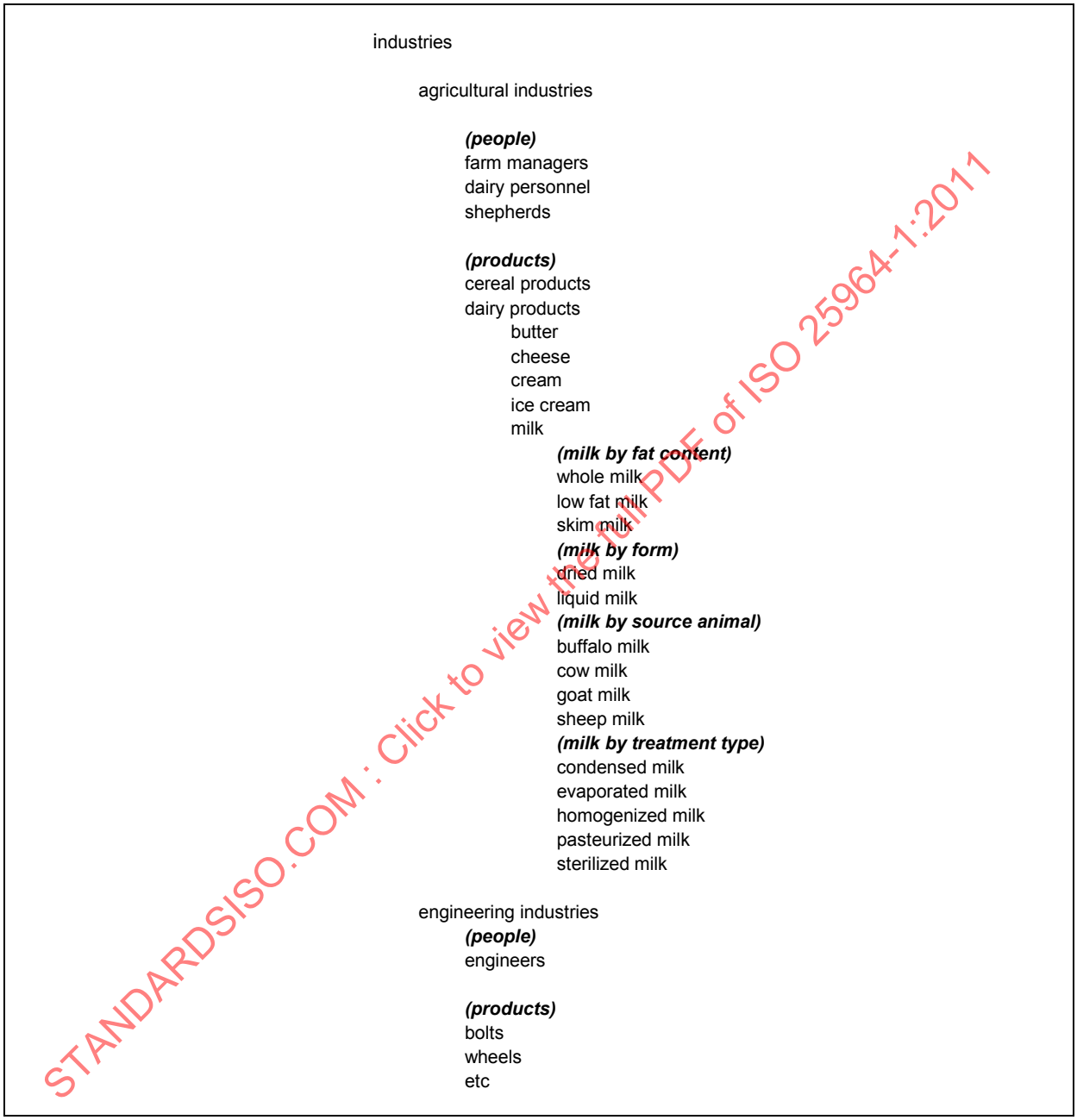


Figure 4 — Partially expanded classified display of an "industries" class, with node labels indicating changes of facet and characteristics of division of arrays

12 Presentation and layout

12.1 General

12.1.1 Use of a database for maintaining presentation information

Within a computer system a thesaurus will normally be held in a database structure (see Clause 15), in which each individual term and each link between a pair of terms is stored once only. This maintains the consistency of the relationships and allows terms to be extracted and displayed in different ways.

12.1.2 Summary of display styles

When presented for human use, on a screen or in printed form, a thesaurus may be displayed in the following ways.

a) A single record is the most elementary form of display, showing the preferred term or non-preferred term itself and (optionally) any or all of the relationships, codes and notes that attach to it.

b) An alphabetical arrangement allows access to concepts from the words in which they are initially expressed by the user. In a printed thesaurus it acts as an index, while in a computer display it can supplement a direct search function.

c) A hierarchical sequence based on BT/NT relationships helps in expanding or refining the concept being indexed or sought.

d) A classified sequence allows browsing in a subject area, drawing attention to related concepts.

e) A graphical display shows terms and their relationships laid out pictorially.

f) A permuted display helps to find words embedded in multi-word terms.

These different arrangements complement each other. Alphabetical access is essential and should be provided by direct search or by an alphabetical sequence as described in b). At least one of c) and d) should be provided to give an overview of the systematic arrangement. Additional information about each concept, such as scope notes and relationships to other concepts, may be shown in any of these sequences. Thesauri can differ widely in their approaches to the relative importance and functions of the different sequences, and also in the arrangement and the kind of relational information provided in each. A graphical presentation can sometimes give a useful overview of a field, but it is complex to create and update and the structure is obscured if it contains too much additional information. A permuted display serves as an index to all the words in all the terms (preferred and non-preferred), using either KWIC (key word in context) or KWOC (key word out of context) format; this is useful in a printed thesaurus, but not usually necessary in a computerized system in which searches for character strings are possible.

12.1.3 Notation and links between sequences

When presented in printed form, the alphabetical sequence should act as an index to show where terms are located in the other sequences, using either line numbers, a system of notation, or, if the hierarchies are not too extensive, by showing the top term of each hierarchy in which the term of interest will be found. To provide links to a graphical display from the alphabetical index it might be necessary to use a system of coordinates. Explicit linking symbols might not be necessary when a thesaurus is displayed on a computer screen, as automatically generated hyperlinks should allow easy switching from one sequence to another, while keeping the focus on the term of interest.

In a classified sequence, notation could, but need not, be expressive of the classified structure and can be designed to allow synthesis of notational symbols to create pre-coordinated class numbers for complex concepts. Any system of notation should provide for the insertion of new concepts at any point in the classified sequence.

12.2 Alternative display styles

12.2.1 General

Many different styles of display are possible, with too many variations to illustrate here. Sections 12.2.2 to 12.2.6 show examples of the basic types of display. To facilitate comparison and contrast, most are illustrated using a common set of terms expressing the concept of "cameras" and related concepts. The content has been selected to include all the basic types of relationships, as well as some optional features such as node labels and notation, which are not so easy to manage. Where notation is present, a simple line-number system has been used. Further examples of displays from published thesauri are included in Annex A.

NOTE Multilingual displays have been omitted from this subclause, for the sake of clarity. Instead they are described in 12.3.

12.2.2 Single record display

When the record for a concept or a preferred term is viewed individually the preferred term should normally be displayed first and the other fields are customarily presented after it in the following sequence:

a) **CC** concept code or notation, locating the term/concept in a hierarchical sequence;

b) **SN** scope note;

c) **UF** references to non-preferred equivalent term(s);

d) **TT** references to top terms;

e) **BT** references to broader terms;

f) **NT** references to narrower terms;

g) **RT** references to related terms;

h) **DEF** definition of the preferred term (see 6.2.3);

i) **HN** history note (see 6.2.4);

j) **SC** subject category or other concept group to which the concept has been assigned.

In this arrangement, the **SN** and **UF** are given near the beginning of the list because they clarify the scope of the concept. They are followed by relationships with other terms. **DEF** and **HN** are given last because they are viewed as administrative fields, used more by the thesaurus editor than the indexer or searcher. As an option, these last fields may be shown after the scope note, bringing all text fields together. However, it is important to ensure that **SN** and **DEF** are not confused when both are present.

The sequence for a non-preferred term is as follows:

— **USE** reference to corresponding preferred term;

— **DEF** definition of the non-preferred term (see 6.2.3);

— **HN** history note (see 6.2.4).

For particular applications the selection and sequence of fields may be varied. For example, the thesaurus editor might wish to include some housekeeping fields (see, for example, 15.2.4, 15.2.9, 15.2.14) in a display for his/her own use.

12.2.3 Alphabetical display

12.2.3.1 In an alphabetical display all terms, whether preferred or non-preferred, are organized as a single alphabetical sequence. Preferred and non-preferred terms should be distinguished typographically whenever possible, so that the non-preferred status of the non-preferred terms is obvious. (In printed media italics may be used for this purpose.) For some purposes a simple list of the terms alone may be sufficient. More commonly the complete record for each term is included, as in 12.2.2.

12.2.3.2 An example of this form of display is shown in Figure 5. Apart from single record displays, this is probably the easiest type of display to construct and reproduce. The most common convention is to show immediate broader and narrower terms only, in which case space is saved but any additional levels of hierarchy are not obvious.

35 mm cameras
CC: H012
BT: film cameras

aqualungs
CC: D002
BT: diving equipment

camera accessories
CC: H002
BT: photographic equipment
NT: flash guns
light meters
tripods
RT: cameras

camera components
CC: H006
BT: cameras and camera components
NT: camera lenses
camera viewfinders

camera lenses
CC: H007
BT: camera components

camera viewfinders
CC: H008
BT: camera components

cameras
CC: H009
BT: cameras and camera components
NT: digital cameras
film cameras
instant picture cameras
plate cameras
reflex cameras
special-purpose cameras
RT: camera accessories
photography

cameras and camera components
CC: H005
BT: photographic equipment
NT: camera components
cameras

colour
CC: A204
BT: optical properties

contrast
CC: A205
BT: optical properties

digital cameras
CC: H010
UF+: *underwater digital cameras*
BT: cameras

divers
CC: D008
BT: people
RT: diving

diving
CC: D000
BT: fields of work
RT: divers
diving equipment

diving equipment
CC: D001
BT: equipment
NT: aqualungs
diving suits
face masks
swimming fins
underwater cameras
RT: diving

diving suits
CC: D003
BT: diving equipment
NT: dry suits
wet suits

dry suits
CC: D004
BT: diving suits

equipment
CC: A001
NT: diving equipment
electrically-powered equipment
fixed equipment
human-powered equipment
photographic equipment
physics equipment
portable equipment

exposure meters
USE: light meters

face masks
CC: D006
BT: diving equipment

fields of work
CC: A300
NT: diving
photography
physics

film cameras
CC: H011
BT: cameras
NT: 35 mm cameras
medium format cameras
miniature cameras

instant picture cameras
CC: H015
SN: Cameras which produce a finished print directly
UF: *Polaroid cameras*®
BT: cameras

people
CC: A100
NT: adults
children
divers
infants
models (people)
photographers
physicists

photographic equipment
CC: H001
BT: equipment
NT: camera accessories
cameras and camera components
RT: photography

photography
CC: H000
BT: fields of work
RT: cameras
photographers
photographic equipment

physical properties
CC: A202
BT: properties
NT: pressure
temperature

physicists
CC: P005
BT: people
RT: physics

physics
CC: P000
BT: fields of work
RT: physicists

Polaroid® cameras
USE: instant picture cameras

pressure
CC: A208
BT: physical properties

properties
CC: A200
NT: physical properties

reflex cameras
CC: H017
SN: Cameras in which the image is reflected on to a glass screen for composing and focusing.
BT: cameras
NT: single lens reflex cameras
twin lens reflex cameras

single lens reflex cameras
CC: H018
UF: *SLR cameras*
BT: reflex cameras

SLR cameras
USE: single lens reflex cameras

special-purpose cameras
CC: H021
BT: cameras
NT: stereo cameras
underwater cameras

stereo cameras
CC: H022
BT: special-purpose cameras

swimming fins
CC: D007
BT: diving equipment

temperature
CC: A209
BT: physical properties

tripods
CC: H004
BT: camera accessories

twin lens reflex cameras
CC: H019
BT: reflex cameras

underwater cameras
CC: D028, H023
UF+: *underwater digital cameras*
BT: special purpose cameras
diving equipment
RT: diving

underwater digital cameras
USE: digital cameras
\+ underwater cameras

Figure 5 — Alphabetical display of thesaurus terms (some terms omitted to save space)

12.2.3.3 Alternatively, it can be useful to indicate more than one level of subordination and/or superordination by different levels of indentation and/or by numbering the levels as shown in the example. If this convention is used to show all the levels of hierarchy for every term, however, it can produce a lengthy display.

EXAMPLE

```
Bovidae
    BT1 ruminants
        BT2 mammals
            BT3 vertebrates
    NT1 buffaloes
    NT1 cattle
        NT2 beef cattle
        NT2 dairy cattle
    NT1 yaks
```

12.2.4 Hierarchical display

A hierarchical display for one language is constructed by indenting each array of narrower terms under its parent broader term. Other relationships (such as equivalence and associative relationships) are not normally shown in this form of display as they would make the structure more difficult to follow. Many hierarchies may result, each headed by a distinct top term (a term with no broader term relationships). For an example, see Annex A, Figure A2.

Alternatively, by applying facet analysis at the top level of the thesaurus, all terms from the same facet may be grouped into a single hierarchy headed by a node label containing the facet name, as shown in Figure 6. (The facet names in Figure 6 are "objects", "fields of work", "people" and "properties".)

When facet analysis is applied in this way, terms are organized into facets according to the basic categories of concept they represent (see 10.2.1 and Clause 11), with no initial regard for the field or fields with which a given concept is usually associated.

Comparing this approach with that described in 12.2.5.2, where the primary organization is by subject and facets are applied within these, primary organization of terms into facets offers certain advantages:

a) the need for a major revision is less likely to occur if a given concept changes the field with which it is usually associated;

b) a higher level of agreement between different documentation centres could be expected.

It is also necessary, however, to note the following disadvantages.

a) Organization by facets tends to scatter the concepts usually associated with a given field or discipline.

b) The basis on which the thesaurus is organized will be generally less self-evident to both indexers and searchers.

c) If the thesaurus covers all disciplines, or a broad range of disciplines, bringing the entities and activities of different disciplines together is unlikely to produce a useful and intelligible sequence of terms. This type of display is, therefore, particularly useful in a thesaurus devoted to a single discipline.

(objects)
equipment
. *<equipment by portability>*
. . fixed equipment
. . portable equipment
. *<equipment by power source>*
. . electrically-powered equipment
. . human-powered equipment
. *<equipment by application>*
. . diving equipment
. . . aqualungs
. . . diving suits
. . . dry suits
. . . wet suits
. . . face masks
. . . swimming fins
. . . underwater cameras
. . photographic equipment
. . . camera accessories
. . . . flash guns
. . . . light meters
. . . . tripods
. . . cameras and camera components
. . . . camera components
. . . . . camera lenses
. . . . . camera view finders
. . . . cameras
. . . . . *<cameras by imaging technique>*
. . . . . digital cameras
. . . . . film cameras
. . . . . *<film cameras by film size>*
. . . . . 35 mm cameras
. . . . . medium format cameras
. . . . . miniature cameras
. . . . . instant picture cameras
. . . . . plate cameras
. . . . . *<cameras by viewing method>*
. . . . . reflex cameras
. . . . . single lens reflex cameras
. . . . . twin lens reflex cameras
. . . . . viewfinder cameras
. . . . . special-purpose cameras
. . . . . stereo cameras
. . . . . underwater cameras

(objects) (cont.)
equipment *(cont.)*
. *<equipment by application> (cont.)*
. . physics equipment
. . . optical instruments
. . . . light meters
. . . . microscopes

(fields of work)
fields of work
. diving
. photography
. physics

(people)
people
. *<people by age>*
. . infants
. . children
. . adults
. *<people by profession>*
. . divers
. . models (people)
. . photographers
. . physicists

(properties)
properties
. physical properties
. . optical properties
. . . colour
. . . contrast
. . . luminance
. . . luminosity
. . pressure
. . temperature

Figure 6 — Hierarchical display of thesaurus terms arranged by facets

12.2.5 Classified display

12.2.5.1 Thesaurus terms arranged by subject fields

The approach of arranging thesaurus terms by disciplines or subject fields is similar to that adopted by the developer of a library classification who begins by organizing the universe of concepts into main classes or disciplines. When applied to a thesaurus, categories of concepts should be grouped initially to reflect the various fields of interest of its users. This technique applies to a thesaurus covering a range of different domains. Concepts that are usually associated with a given field, such as "art", are brought together, and are also effectively separated from those belonging to different areas of interest, such as "economics" or "physics".

This approach can also be applied to a thesaurus that is restricted to a single field. A thesaurus devoted to "medicine", for example, might be organized initially into subfields such as "surgery", "physiology", "anatomy", etc. To complement the core concepts, concepts from different disciplines may also be used, for example "management", "law" and "data processing", which should be separated from the medical terms.

The subject approach has the advantage of organizing concepts into groups that generally correspond to the modes of thinking of the users. One difficulty is that some concepts are used habitually in several different fields, for example, if a thesaurus covers forestry as well as agronomy, there will be many concepts, such as planting, which occur in both. Such concepts should be either repeated in each field in which they occur or given in a separate list of general concepts not limited to a specific field.

Grouping concepts by subject field, with sibling terms arranged either alphabetically or systematically, and with indentation of some under others, produces displays that look like hierarchies but in which the subordinate terms might not follow the rules for hierarchical relationships described in 10.2. Node labels giving the facet names should, therefore, be inserted into such displays to indicate where different facets occur and where the relationships are not hierarchical, as discussed in Clause 11.

12.2.5.2 Thesaurus terms arranged by facets within subject fields

In practice, the different approaches described in 12.2.4 and 12.2.5.1 are frequently combined, for example when a thesaurus organized primarily into subject fields is further subdivided by facets.

Figure 7 illustrates the combined approach using the example of a thesaurus that takes in the fields of diving, photography and physics. Only the preferred terms relevant to a particular field are shown within that field. Some terms common to all three fields are presented in a section of general terms. After the primary organization by subject, the next subdivision is by facet. In Figure 7, the facets "objects", "people" and "properties" have been applied to each subject section in turn.

In printed displays, it is often necessary to limit the total size of the compilation by choosing one principal location for each term. The term may be repeated in other locations but space is saved by avoiding repetition of all its narrower terms. The user is expected to refer to the principal location to find the complete hierarchy. Thus, in Figure 7, the "principal location" of "properties" is at A200, where "chemical properties", "physical properties" and all their narrower terms are also displayed. In the diving section, the only physical properties to be mentioned are "pressure" and "temperature". In the photography section, only the optical properties are fully spelt out.

Figure 7 also illustrates the use of notation for locating terms in a classified display. In this example, a term has only one notation, corresponding to its principal location. Where the term is repeated in other locations, it is followed by the notation in brackets, helping the user to refer to the principal location. Some thesauri allow a term to have subsidiary notations, so that every repetition can easily be located.

The designation of a "principal location" is convenient for restricting the size of a printed thesaurus, but adds extra steps to the process of thoroughly researching a concept. Computer displays can make the look-up process much easier if they allow users to expand or collapse portions of hierarchy, showing as much detail as users require, only when they want to see it.

A000 general terms
(objects)
A001 . equipment
. . *<equipment by portability>*
A002 . . . fixed equipment
A003 . . . portable equipment
. . *<equipment by power source>*
A004 . . . electrically-powered equipment
A005 . . . human-powered equipment
. . *<equipment by application>*
. . . diving equipment (D001)
. . . photographic equipment (H001)
. . . physics equipment (P001)
(people)
A100 . people
. . *<people by age>*
A101 . . . infants
A102 . . . children
A103 . . . adults
. . *<people by profession>*
. . . divers (D008)
. . . models (people) (H025)
. . . photographers (H026)
. . . physicists (P005)
(properties)
A200 . properties
A201 . . chemical properties
A202 . . physical properties
A203 . . . optical properties
A204 . . . . colour
A205 . . . . contrast
A206 . . . . luminance
A207 . . . . luminosity
A208 . . . pressure
A209 . . . temperature

(fields of work)
A300 . fields of work
. . diving (D000)
. . photography (H000)
. . physics (P000)

D000 diving
(objects)
. equipment (A001)
D001 . . diving equipment
D002 . . . aqualungs
D003 . . . diving suits
D004 . . . . dry suits
D005 . . . . wet suits
D006 . . . face masks
D007 . . . swimming fins
. . underwater cameras (H023)
(people)
. people (A100)
D008 . . divers
(properties)
. properties (A200)
. . physical properties (A202)
. . . pressure (A208)
. . . temperature (A209)

H000 photography
(objects)
. equipment (A001)
H001 . . photographic equipment
H002 . . . camera accessories
H003 . . . . flash guns
. . . . light meters (P003)
H004 . . . . tripods
H005 . . . cameras and camera components
H006 . . . . camera components
H007 . . . . . camera lenses
H008 . . . . . camera viewfinders
H009 . . . . cameras
. . . . . *<cameras by imaging technique>*
H010 . . . . . digital cameras
H011 . . . . . film cameras
. . . . . *<film cameras by film size>*
H012 . . . . . 35 mm cameras
H013 . . . . . medium format cameras
H014 . . . . . miniature cameras
H015 . . . . . instant picture cameras
H016 . . . . . plate cameras
. . . . . *<cameras by viewing method>*
H017 . . . . . reflex cameras
H018 . . . . . single lens reflex cameras
H019 . . . . . twin lens reflex cameras
H020 . . . . . viewfinder cameras
H021 . . . . special-purpose cameras
H022 . . . . . stereo cameras
H023 . . . . . underwater cameras
H024 . . . . . high pressure underwater cameras

(people)
. people (A100)
H025 . . models (people)
H026 . . photographers
(properties)
. properties (A200)
. . physical properties (A202)
. . . optical properties (A203)
. . . . colour (A204)
. . . . contrast (A205)
. . . . luminance (A206)
. . . . luminosity (A207)

P000 physics
(objects)
. equipment (A100)
P001 . . physics equipment
P002 . . . optical instruments
P003 . . . . light meters
P004 . . . . microscopes
(people)
. people (A100)
P005 . . physicists
(properties)
. properties (A200)
. . physical properties (A202)

Figure 7 — Classified display of thesaurus terms arranged by facets within subject fields

12.2.5.3 Expanded classified display

It is possible to add to the type of display shown in Figure 7 all of the data described in the single record display, as in 12.2.2, in order to present each term in every one of its locations. Although Figure 8 does not go to this length, it illustrates the inclusion of additional information such as scope notes, non-preferred terms and related terms, but not any cross-language equivalents.

NOTE Compare, for example, underwater cameras at H023.

The inclusion of extra data adds to length. To counteract at least some of this, Figure 8 adopts a slightly different convention regarding repetition of hierarchies. Comparing Figure 8 with Figure 7, it can be seen that the mention of "pressure" and "temperature" in the diving section of Figure 8 does not include their broader terms "physical properties" and "properties" as shown at the corresponding location in Figure 7. To find these, the user is obliged to follow the notation links to A208 and A209, where the full hierarchies are visible. Apart from saving some space, this approach avoids giving a false impression of the scope of a broad term such as properties or physical properties, when appearing in its secondary locations.

To conclude the comparison of the merits of Figure 6, Figure 7 and Figure 8, it should be clarified that no one of these is being recommended as the best for all situations. Each of the approaches has advantages and disadvantages, the importance of which varies according to the context of application. Further variations such as concept groups are also possible, as described in 15.2.18 (see also Annex A, Figures A2, A4, A7, A9, A10, A13, A16.) However, it is essential that at least one form of systematic display, hierarchical or classified, should be available to users in addition to the customary alphabetical display. (See also 12.3.4 for inclusion of multilingual issues).

A000 general terms

<*objects*>
A001 . equipment
. . <*equipment by portability*>
A002 . . . fixed equipment
A003 . . . portable equipment
. . <*equipment by power source*>
A004 . . . electrically-powered equipment
A005 . . . human-powered equipment
. . <*equipment by application*>
. . . diving equipment [D001]
. . . photographic equipment [H001]
. . . physics equipment [P001]
<*people*>
A100 . people
. . <*people by age*>
A101 . . . infants
A102 . . . children
A103 . . . adults
. . <*people by profession*>
. . . divers [D008]
. . . models (people) [H025]
. . . photographers [H026]
. . . physicists [P005]
<*properties*>
A200 . properties
A201 . . chemical properties
A202 . . physical properties
A203 . . . optical properties
A204 . . . . colour
A205 . . . . contrast
A206 . . . . luminance
RT: light meters [P003]
A207 . . . . luminosity
A208 . . . pressure
A209 . . . temperature
<*fields of work*>
A300 . fields of work
. . diving [D000]
. . photography [H000]
. . physics [P000]

D000 diving

<*objects*>
D001 . diving equipment
BT: equipment [A001]
D002 . . aqualungs
D003 . . diving suits
D004 . . . dry suits
D005 . . . wet suits
D006 . . face masks
D007 . . swimming fins
. . underwater cameras [H023]
<*people*>
D008 . divers
BT: people [A100]
<*properties*>
. pressure [A208]
. temperature [A209]

H000 photography

<*objects*>
H001 . photographic equipment
BT: equipment [A001]
H002 . . camera accessories
H003 . . . flash guns
. . . light meters [P003]
H004 . . . tripods
H005 . . cameras and camera components

photography (cont.)

<*objects*>(cont.)
. photographic equipment*(cont.)*
. . cameras and camera components *(cont.)*
H006 . . . camera components
H007 . . . . camera lenses
H008 . . . . camera viewfinders
H009 . . . cameras
. . . . <*cameras by imaging technique*>
H010 . . . . . digital cameras
UF+: *underwater digital cameras*
H011 . . . . . film cameras
. . . . . <*film cameras by film size*>
H012 . . . . . 35 mm cameras
H013 . . . . . medium format cameras
H014 . . . . . miniature cameras
H015 . . . . . instant picture cameras
SN: Cameras which produce a finished print directly
UF: *Polaroid® cameras*
H016 . . . . . plate cameras
. . . . <cameras by viewing method>
H017 . . . . . reflex cameras
SN: Cameras in which the image is reflected on to a glass screen for composing and focusing
H018 . . . . . single lens reflex cameras
UF: *SLR cameras*
H019 . . . . . twin lens reflex cameras
H020 . . . . . viewfinder cameras
H021 . . . . special-purpose cameras
H022 . . . . . stereo cameras
H023 . . . . . underwater cameras
SN: cameras for taking pictures under water
UF+: *underwater digital cameras*
BT: diving equipment [D001]
H024 . . . . . high pressure underwater cameras
<*people*>
. people [A100]
H025 . . models (people)
H026 . . photographers
<properties>
. optical properties [A203]
. . colour [A204]
. . contrast [A205]
. . luminance [A206]
. . luminosity [A207]

P000 physics

<*objects*>
P001 . physics equipment
BT: equipment [A100]
P002 . optical instruments
P003 . . light meters
UF: *exposure meters*
RT: luminance [A206]
P004 . . microscopes
<*people*>
P005 . physicists
BT: people [A100]
<*properties*>
. physical properties [A202]
. . optical properties [A203]
. . . colour [A204]
. . . contrast [A205]
. . . luminance [A206]
. . . luminosity [A207]
. . pressure [A208]
. . temperature [A209]

Figure 8 — Classified display of thesaurus terms arranged by facets within subject fields, with additional information such as SN, UF and RT

12.2.6 Displaying polyhierarchical relationships

12.2.6.1 The standard treatment of polyhierarchical relationships (see 10.2.5) is demonstrated in Figure 5, in which the term “underwater cameras” has the two broader terms "special purpose cameras" and "diving equipment". Therefore, it appears at two positions in the equipment facet of Figure 6. If it had narrower terms as well, these would be shown in both positions.

12.2.6.2 A different approach may be adopted in the classified or hierarchical section of a printed thesaurus in which space is strictly limited and an extensive entry is needed for the polyhierarchical term. It might have a large hierarchy of narrower terms, and possibly also a long scope note and/or several non-preferred synonyms and related terms. As in Figure 6 and Figure 7, one "principal location" might be chosen for the polyhierarchical term, and in other locations a less complete display might be shown. See, for example, the treatment of "underwater cameras" in Figure 8.

12.2.6.3 Computer displays should offer users the facility to see all the relationships for a polyhierarchical term, wherever it occurs.

12.3 Presentation and layout of multilingual thesauri

12.3.1 General

In addition to the general recommendations applying to all thesauri (see 12.1 and 12.2), two important requirements should be taken into account when dealing with presentation and layout of multilingual thesauri:

a) To confer equal status on all languages of the multilingual thesaurus, the alphabetical sequence conventionally established for each language should be respected (see 12.4.3). This applies not only to alphabetical presentations of thesauri, but also to alphabetical sequences found in preferred term records.

b) Cross-language equivalents should be available for preferred terms in each language version.

In addition to the examples shown in 12.3, Annex A shows several further examples taken from published thesauri.

12.3.2 Single-record display

The user interface should allow selection of one of the included languages as the basis for single record display. The display described in 12.2.2 should be augmented to include all cross-language equivalents. Each language is identified by an alpha-2 language code, as stipulated by ISO 639-1 (see Figure 9). If an appropriate code is not found in ISO 639-1, an alpha-3 code from ISO 639-2 may be used. These codes may be extended where necessary with the additional codes described in RFC 4646[45] and listed in the IANA subtag registry[35].

NOTE The colons shown in Figures 9, 10 and 11 are optional.

pesticides	
es:	plaguicidas
CC:	FF120
SN:	Excludes growth regulators and repellents
UF:	*fumigants*
TT:	agrochemicals
BT:	agrochemicals
NT:	fungicides
	herbicides
	insecticides
RT:	pests
DEF:	A substance for destroying pests
HN:	Added 1975-04-01

Figure 9 — Single record display in a bilingual thesaurus (English – Spanish)

A similar, complete, single-record display should be available for each term in each language version of the multilingual thesaurus. Thus if the editor chooses the second language available in this particular bilingual Spanish-English thesaurus, the display shown in Figure 10 should be visible.

plaguicidas	
en:	pesticides
CC:	FF120
SN:	Excluye repelentes y reguladores del crecimiento
UF:	*fumigantes*
TT:	productos químicos agrícolas
BT:	productos químicos agrícolas
NT:	fungicidas
	herbicidas
	insecticidas
RT:	plagas
DEF:	Agente que combate las plagas del campo
HN:	Agregado 1975-04-01

Figure 10 — Single record display in a bilingual thesaurus (Spanish - English)

Except where equivalences between non-preferred terms have been established (see 9.5), the single record display for a non-preferred term in a multilingual thesaurus is identical to that in a monolingual thesaurus (see 12.2.2).

12.3.3 Alphabetical display

There should be a distinct alphabetical sequence for each language of a multilingual thesaurus. Integration of multiple languages in a single alphabetic sequence is not recommended.

The preferred term records in the alphabetical display for a given language should be augmented with the corresponding cross-language equivalents in each language of the multilingual thesaurus. Records for non-preferred terms do not require such augmentation. Figure 11 shows an extract from a bilingual display in this style. A corresponding display should be prepared for each of the languages of the thesaurus.

mosquitoes
  es: mosquitos
  CC: HH120
  BT: pests

pesticides
  es: plaguicidas
  CC: FF120
  SN: Excludes growth regulators and repellents
  UF: *fumigants*
  BT: agrochemicals
  NT: fungicides
  herbicides
  insecticides
  RT: pests

pests
  es: plagas
  CC: HH100
  NT: locusts
  mosquitoes
  slugs
  snails
  RT: pesticides

pigs
  es: cerdos
  CC: HB130
  UF: *hogs*
  porkers
  sows
  BT: livestock

plant products
  es: productos de origen vegetal
  CC: PD100
  NT: cereals
  fruits
  spices
  vegetables
  RT: plants

plants
  es: plantas
  CC: HM100
  RT: plant products

porkers
  USE: pigs

poultry
  es: aves de corral
  CC: HB140
  BT: livestock
  NT: chickens
  ducks
  geese
  turkeys
  RT: eggs

sheep
  es: ovinos
  CC: HB190
  BT: livestock
  RT: wool

Figure 11 — Alphabetical display for the English version of a bilingual thesaurus

12.3.4 Systematic displays

It can be useful to see the structure of the thesaurus laid out in parallel in each of the language versions. Either a hierarchical display, as in Figure 6, or a classified display, as in Figure 7, may be chosen. These separate monolingual displays, if they are prepared for each language, might be adequate when viewed side by side. Alternatively, the language versions may be presented in columns on the same page, so that the correspondence is plainly visible. Figure 12 shows an excerpt from a hierarchical display in this style.

Figure 12 shows notation in a central column for ease of reading across. Notation is optional, but more useful in a multilingual thesaurus than a monolingual one, because it is independent of language. It should be noted, however, that alphabetical ordering of siblings in more than one language simultaneously is usually impossible to achieve, thus giving the impression that one language is dominant. To avoid this problem, some other sequence of siblings may be used throughout, in both languages.

In a multilingual thesaurus of more than two or three languages, parallel hierarchical displays become cumbersome. An alternative option for online displays is to use a split screen approach whereby one side of the screen shows the hierarchical display in one language only, while the other side of the screen shows equivalents in all the other languages for the preferred term selected in the hierarchy. Figure A.11 in Annex A illustrates this approach. Figure A.7, although monolingual, is also helpful for visualizing what can be done.

<materials>	FF100	<materiales>
. agrochemicals	FF110	. productos químicos agrícolas
. . fertilizers	FF120	. . fertilizantes
. . pesticides	FF130	. . plaguicidas
. . . fungicides	FF140	. . . fungicidas
. . . herbicides	FF150	. . . herbicidas
. . . insecticides	FM100	. . . insecticidas
. feeds	FP100	. piensos
. fuel		. combustibles
<organisms>	HB100	<organismos>
. livestock	HB110	. ganado
. . cattle	HB120	. . ganado bovino
. . goats	HB130	. . caprinos
. . pigs	HB140	. . cerdos
. . poultry	HB150	. . aves de corral
. . . chickens	HB160	. . . pollos
. . . ducks	HB170	. . . patos
. . . geese	HB180	. . . gansos
. . . turkeys	HB190	. . . pavos
. . sheep	HH100	. . ovinos
. pests	HH110	. plagas
. . locusts	HH120	. . langostas de tierra
. . mosquitoes	HH130	. . mosquitos
. . slugs	HH140	. . babosas
. . snails	HM100	. . caracoles
. plants		. plantas
<products>	PC100	<productos>
. animal products	PC200	. productos animals
. . dairy products	PC210	. . productos lácteos
. . . butter	PC220	. . . mantequilla
. . . cheese	PC230	. . . queso
. . . cream	PC240	. . . crema
. . . milk	PC250	. . . leche
. . eggs	PC260	. . huevos
. . leather	PC270	. . cuero
. . meat	PC300	. . carne
. . wool	PD100	. . lana
. plant products	PD110	. productos de origen vegetal
. . cereals	PD120	. . cereales
. . fruits	PD130	. . frutas
. . spices	PD140	. . especias
. . vegetables		. . hortalizas

Figure 12 — Hierarchical display for a bilingual thesaurus

12.3.5 Correspondence tables for multilingual thesauri

For some purposes it is sufficient to display only the equivalences between the languages in a correspondence table, without any other relationships. Figures 13 and 14 show a convenient layout. A separate table should be prepared for each of the languages of a multilingual thesaurus, each table applying the alphabetical filing conventions of the language used as source language.

English	**Spanish**
agrochemicals	productos químicos agrícolas
agricultural chemicals see: agrochemicals	
animal products	productos animales
butter	mantequilla
cattle	ganado bovino
cereals	cereales
cheese	queso
chickens	pollos
cream	crema
dairy products	productos lácteos
ducks	patos
eggs	huevos
feeds	piensos
fertilizers	fertilizantes
fruits	frutas
fuel	combustibles
fumigants see: pesticides	
fungicides	fungicidas
geese	gansos
goats	caprinos
hens see: chickens	
herbicides	herbicidas
hogs see: pigs	
insecticides	insecticidas
leather	cuero
livestock	ganado
locusts	langostas de tierra
meat	carne
milk	leche
mosquitoes	mosquitos
pesticides	plaguicidas
pests	plagas
pigs	cerdos
plant products	productos de origen vegetal
plants	plantas
porkers see: pigs	
poultry	aves de corral
sheep	ovinos
slugs	babosas
snails	caracoles
sows see: pigs	
spices	especias
turkeys	pavos
vegetables	hortalizas
wool	lana

Figure 13 — Correspondence table for a bilingual thesaurus (English – Spanish)

Español	Ingles
aves de corral	poultry
babosas	slugs
caprinos	goats
caracoles	snails
carne	meat
cerdos	pigs
cereals	cereals
chanchos véase: cerdos	
combustibles	fuel
crema	cream
cuero	leather
especias	spices
fertilizantes	fertilizers
frutas	fruits
fumigantes véase: plaguicidas	
fungicidas	fungicides
Ganado	livestock
ganado bovino	cattle
ganado porcino véase: cerdos	
gansos	geese
herbicidas	herbicides
hortalizas	vegetables
huevos	eggs
insecticidas	insecticides
lana	wool
langostas de tierra	locusts
leche	milk
mantequilla	butter
mosquitos	mosquitoes
ovinos	sheep
patos	ducks
pavos	turkeys
piensos	feeds
plagas	pests
plaguicidas	pesticides
plantas	plants
pollos	chickens
productos animales	animal products
productos de origen vegetal	plant products
productos lácteos	dairy products
productos químicos agrícolas	agrochemicals
puercos véase: cerdos	
queso	cheese

Figure 14 — Correspondence table for a bilingual thesaurus (Spanish – English)

12.4 Language and character encoding issues

12.4.1 General

The different languages used in a multilingual vocabulary sometimes use different writing systems. Even languages that share many characters (such as languages using the Latin alphabet) sometimes use different diacritics or other symbols in conjunction with some of those characters. Transliteration schemes are also

used to transpose characters in these languages into other writing systems, where there are limitations of printing or display, or in the ability of users to read unfamiliar scripts.

Use of the Universal Character Set (UCS) is recommended, in order to provide for as many languages and scripts as possible. The UCS is defined jointly by the Unicode Standard and ISO/IEC 10646.

The UCS has been specified in other standards (see References [57] to [59]), and has been widely supported by computer manufacturers and software developers. The use of these standards for character encoding promotes interoperability with other systems. In the case that rare characters are found beyond the coverage of the Universal Character Set, national or industry standards should be consulted and appropriately incorporated.

Care should be taken to ensure that all components of a vocabulary management system are compatible and properly configured with the particular languages and character sets. These components include not only the vocabulary management software but also computer hardware for data entry (e.g. keyboards, input software), computer operating systems, and fonts and formats for display purposes (for print publications and computer-based displays). Special consideration should also be given to display issues, filing orders and normalization required for searching and index creation.

12.4.2 Display issues

In each language, characters should appear consistently throughout the vocabulary in a form acceptable to speakers of that language, including diacritics where appropriate. Although German speakers sometimes find it acceptable to convert "ä" to "ae" and "ö" to "oe" etc., it is preferable that the form in which characters appear on the printed page or on the computer screen is the form most familiar to users. Languages that are written from right to left, such as Arabic and Hebrew, usually need special consideration to ensure that text appears correctly displayed and aligned.

12.4.3 Filing orders

Different languages (and even different contexts within the same language) sometimes conventionally require the use of different filing orders, also known as sorting orders. For each language, it might be necessary to specify both the characters and diacritics that are ignored for filing purposes and the sequence in which characters should file. For example:

a) the difference between upper and lower case letters is ignored in sorting for English (and in most other languages in the Latin, Greek, Cyrillic and Georgian scripts);

b) in Spanish, the letter ñ files after the letter n, and before the letter o;

c) in French, all diacritics are ignored in filing and the letter ñ interfiles with the letter n;

d) in Czech, ch is treated as a single letter coming after the letter h and before i;

e) in Croatian, dž, lj and nj are also treated as single letters coming after d, l and n, and before đ, m and o respectively;

f) several alternatives exist for sorting of ideographic characters, as used in Chinese, Japanese and Korean. In addition, simplified and traditional versions of ideographic characters exist, although the simplified and traditional forms might file identically (and are coded identically in ISO/IEC 10646);

g) numerical strings are sometimes filed as text, e.g. 1, 10, 102, 11, 120, 2, but other times as numbers, e.g. 1, 2, 10, 11, 102, 120.

Refer to References [11] and [3] for basic guidance on filing order suitable for indexes and listings for monolingual applications.

Filing orders in multilingual contexts are defined by ISO/IEC 14651, and also in the equivalent *Unicode Collation Algorithm*. However, users might also find it useful to take account of practice in specific languages, which is documented (for languages that use the Latin script) in ISO 12199.

Computer software for multilingual vocabularies should be capable of producing conventional filing sequences for each of the different languages in order to facilitate user consultation. If the conventional order cannot be supported, text accompanying each language version of the vocabulary should explain clearly the filing order used.

If displays are required as shown in Figure 12, in which visual correspondence across languages is important, any attempt to use alphabetical order for the sequence of siblings produces unsatisfactory results for at least one of the languages. Figure 12 illustrates this problem. To avoid giving any language preference, it is necessary either to:

— produce multiple outputs, one for each language, in which the sequence of all the others is driven by correspondence to the first one; or

— choose an underlying systematic sequence (e.g. from smallest to largest or from south to north), applying across all the languages.

12.4.4 Normalization for information retrieval

Searching, either within a vocabulary management system or within an information retrieval system using terms taken from a controlled vocabulary, involves matching a string of characters entered by a searcher (search term) with a string of characters extracted from metadata and stored into an index (index term). The search term and the index term sometimes differ in ways that are not considered to be significant, including encoding and the use of upper or lower case characters, diacritics or punctuation. Normalization is the process of converting search and index terms into canonical forms that facilitate string matching and thereby enhance information retrieval by increasing recall.

EXAMPLE

If a vocabulary management system contains, as a French term, the name of the organization Alliance coopérative internationale, normalization rules applied during the index building process might convert all characters to upper case, and convert accented characters to unaccented equivalents, so that the index contains the string ALLIANCE COOPERATIVE INTERNATIONALE. Similar rules apply when a user enters a search string so that this term record is retrieved by any of the following search strings:

alliance cooperative internationale

Alliance Cooperative Internationale

Alliance Coopérative Internationale

ALLIANCE COOPERATIVE INTERNATIONALE

ALLIANCE COOPÉRATIVE INTERNATIONALE

Regardless of the form the user enters, the two normalized canonical forms match, and the user successfully retrieves the record for the term in the controlled vocabulary.

The above example presents a simple scenario, within a single vocabulary management system. However, normalization rules can also affect the use of spaces (e.g. in place names or pairs of words such as data bases and databases), punctuation characters (particularly hyphens as in words such as fish-breeding), the way that words are extracted, and the use of special symbols such as & (ampersand). This is described further in ISO/IEC 14651 and in the equivalent *Unicode Collation Algorithm*.

In rare cases, terms for different concepts might have the same normalized form (e.g. AIDS and aids) and should be treated within the vocabulary as homographs (see 6.2.2). Rules for normalization should be made clear to the vocabulary user, and spelling and punctuation rules for the form of terms should take into account the normalization supported by the vocabulary management system.

12.4.5 Identification of dialects and/or alternative scripts

If dialects are treated as separate languages, or if a transliteration scheme is used in conjunction with a particular language, the dialect or transliteration scheme should be identified in the tags used in multilingual displays. The alpha-2 codes from ISO 639-2 should be supplemented with the additional codes described in RFC 4646[45] and listed in the IANA subtag registry[35]. These allow the specification of languages, scripts, regions and other variants, including transliteration. Thus text transliterated from Mandarin Chinese into latin characters using the Pinyin transliteration scheme should be encoded as zh-Latn-pinyin, while English as used in the United States should be encoded as en-US.

13 Managing thesaurus construction and maintenance

13.1 Planning a thesaurus

13.1.1 Determining objectives

Building a thesaurus is a labour-intensive job, requiring commitment for many years if it is to prove worthwhile. Therefore, before significant time and money are spent, objectives should be clarified based on considerations such as the following:

a) what the thesaurus is to be used for, and by whom;

b) whether it will be limited by the constraints of existing software with which it is to be used;

c) how knowledgeable the users will be in the subject area of the thesaurus and in thesaurus use.

Different stakeholders (such as senior managerial staff, IT support staff and the prospective users) have different expectations about what a thesaurus is and how it will perform. A statement of objectives should be agreed by all parties as a key step in procuring effective teamwork.

13.1.2 Determining thesaurus features

When knowledge of objectives and of the planned uses and user community has been obtained, the scope and style of the thesaurus should be defined.

The following features should be determined:

a) whether the thesaurus will be used in print or electronic format, or both;

b) the style of display required;

c) whether special formats are needed for integration with other systems, such as searching or indexing systems;

d) what format is envisaged for updates, and how frequently these will be required.

Decisions should be taken on optional features, such as whether to subdivide the hierarchical relationship as described in 10.2.1, whether to introduce customized relationships (see 10.4), whether to show node labels in output displays and whether it is important to store term definitions, sources, and date information.

At a more detailed level, decisions should be taken on the conventions to be used for singular/plural, spelling and capitalization. It might be useful to consider practical constraints such as the maximum length of term that will be permitted, any diacritics or other special characters needed, and whether or not there needs to be any limit to the number of levels of hierarchy.

13.1.3 Taking stock of resources

13.1.3.1 General

During the planning phases of a thesaurus project, the key resources, such as people, funding, software tools and vocabulary resources should be determined. If the thesaurus is multilingual, the budget should allow for the extra cost of communication overheads, as well as the cost of any specialized software that might be required.

13.1.3.2 Human resources

Human resources should include the following:

a) a champion at the senior management level, and preferably corporate commitment to the thesaurus project;

b) interest and support from users, particularly from any terminology experts in one or more areas of the subject field;

c) information professionals knowledgeable in the subject area, with aptitude and preferably training in thesaurus work, to take on the main task of thesaurus construction and maintenance;

d) support from IT professionals who can help with installation/adaptation of thesaurus management software, any programming required to integrate thesaurus outputs with the systems in which the thesaurus will be applied, and the day-to-day running of normal desktop tools.

If more than one thesaurus editor is involved, careful planning should be undertaken to avoid simultaneous, potentially conflicting, inputs to the same portion of the thesaurus.

If the thesaurus is multilingual, the people who undertake construction and maintenance should ideally have all of the following attributes:

— a good understanding of each of the natural languages involved;

— a good knowledge of the subject area of the vocabulary; and

— a good understanding of the difference between translation and identification of terminological equivalents for information retrieval purposes.

Such people are scarce and often located in different countries. For this reason, it is important to maintain frequent, effective and efficient communication between all team members. Agreement on procedures and formats for unambiguous communication of structured data should be sought.

13.1.3.3 Vocabulary resources

Existing vocabulary resources should be researched before construction begins. Any or all of the following might prove useful:

a) Existing thesauri or classification schemes with a scope that partially or wholly matches the requirement. In extreme cases, it might be cost-effective simply to adopt an existing thesaurus unchanged. More commonly, portions of an existing vocabulary can be taken and/or adapted, subject to copyright. Existing works are a source of ideas for terms, or structure, or both.

b) Collections of terminology or frequently asked queries that colleagues might have assembled in the course of their work.

c) The indexes of existing databases or other relevant reference tools, showing useful entry points to the subject matter. If an existing relevant database has been indexed with a controlled vocabulary, it might be possible to obtain from it a report of all the index terms used, and their frequency of usage.

d) The transaction log of a relevant website, showing the search terms most frequently employed by users.

e) Standard reference works such as dictionaries, terminologies, nomenclatures and encyclopaedias are useful for verification of a term's meaning and validation of hierarchical structures, rather than as a guide to selection of terms.

13.1.4 Establishing responsibilities

At an early stage an editor should be appointed to take responsibility for managing the project through all of its stages, possibly including continuing maintenance as well as compilation of the first edition. Depending on the size of the project, the editor may be supported by an editorial team, and possibly by an editorial review board. There are some advantages in combining thesaurus editorial responsibility with quality control of the subject indexing of resources using the thesaurus.

13.1.5 Choosing thesaurus management software

Although it is possible to compile and produce a thesaurus entirely without software support, this procedure is not recommended. Purpose-built thesaurus management software should be used for efficiency and to avoid clerical errors.

When choosing software, a list of requirements should be drawn up based on the features outlined in Clause 14, and any additional requirements. Existing software products on the market should be evaluated. If none of them meets all of the requirements, some compromises might be necessary. Alternatively, one of the existing software products could be customized, or a new one developed.

13.2 Early stages of compilation

13.2.1 General: how and when to start

Assuming that the thesaurus is needed to support the indexing and searching of a database or other resource, the thesaurus should be completed before database compilation begins in order to ensure that indexing is conducted systematically. However, it is difficult to decide which terms will be needed in the thesaurus until a substantial part of the database has been built. To avoid a later need for re-indexing, at least part of the thesaurus should be prepared in advance, but the need for substantial continuing work on the thesaurus should be anticipated.

Software tools are available that analyze a corpus of text in the relevant subject field(s) and extract significant words or phrases, together with their frequency of use. Some packages also organize the terms into a structured format. While an automatically created thesaurus is unlikely to match the quality of one prepared intellectually, as a source of ideas it can speed up the intellectual process.

13.2.2 Collecting terms and concepts

As stressed in 5.1, the terms appearing in a thesaurus are only representations of the underlying concepts on which the thesaurus is based. During construction of the thesaurus, the overall aim is to collect and organize concepts, but in practice this is achieved by collecting terms and arranging them so that they adequately represent the required concepts.

Prime sources of terms are listed in a) to d) of 13.1.3.3. The full text of literature central to the required scope is also useful. As described in 13.2.1, automatic techniques can be used to extract key words and phrases from the literature. While it is relatively easy to study the material to be input, at least as much emphasis should be given to any indications of what it is that users want to search for, such as lists of queries registered at an enquiry desk. The users themselves may be invited to suggest useful terms and concepts.

At this early stage, it is not necessary to make decisions as to the best representation of a required concept. The initial task is to collect all the terms available, noting the source of each and the frequency of occurrence.

13.2.3 Analysing terms

The terms collected should be sorted into systematic order before they can be studied properly and entered into the thesaurus. If the thesaurus management software does not support this task, standard desktop tools such as spreadsheets can be useful for this preliminary sorting out. The process is made easy and efficient if the terms are obtained from a machine-readable source. Once the collected terms are in a spreadsheet, together with their source and frequency, they should be sorted into subject areas or facets (see Clause 11) so that like terms with all their variants and synonyms are brought together. Although the classification or facet structure employed at this stage might be rudimentary, and need not determine the eventual structure of the thesaurus, it can help in developing ideas for the latter. The main product of this analytical stage should be a list or set of small lists of terms, grouped by subject and/or facet, with indications of the source and frequency of use of each term. Armed with these data, the editor is then able to begin systematic construction.

13.3 Construction

13.3.1 Input

Terms should be handled in subject/facet groups or hierarchies. Groups of synonyms and near-synonyms should be considered together and preferences and equivalence relationships established. At this stage, the terms and their hierarchical relationships should be selected carefully, taking into account potential duplications, overlaps, omissions and the degree of specificity required. Scope notes may also be considered and entered at the same time. If desired, source and definition may also be entered. It is best that the date of input and term identifier are added automatically by the software, rather than by manual keying.

If the initial stocktaking exercise has revealed an existing thesaurus of which a substantial portion should be adopted, it might be possible to import this directly into the chosen thesaurus management software. This procedure can save time, particularly if the software permits import of structured data with scope notes and relationships as well as the terms. However, it is essential that the permission of the copyright holder is obtained.

Unstructured lists of terms can also be imported directly into some software products. This does not save as much time since relationships still need to be entered and careful inspection is needed to ensure that spelling inconsistencies are not introduced accidentally.

13.3.2 Construction sequence

When working with hierarchies, the question arises of whether to build from the top downwards or from the bottom upwards. A view of the top layers is useful in ensuring that the thesaurus has the appropriate coverage and balance to meet the desired scope. It is also useful when inviting contributions from experts and other users to be able to place their advice in the context of the overall structure.

On the other hand, work at the more specific level tends to throw up anomalies that can affect structure at the higher levels. The availability of low-level portions of hierarchy, revealing many concepts that need to be accommodated somewhere, tends to inform development of a sound superstructure.

Both strategies are useful and should be pursued in combination. A sketch of the high-level structure should be prepared before work proceeds on lower-level sections of hierarchy, one at a time. As the work proceeds, the high-level sketch should be modified to meet emerging needs.

It is efficient to work with groups of hierarchically clustered terms, entering them together with equivalence and hierarchical relationships. However, associative relationships cannot be entered as easily at this stage because the terms at the other end of the association are usually in another hierarchy, still to be developed. Associative relationships are usually entered at a later stage when the bulk of the terms are present in the system.

Whichever sequence is followed, quality control should be undertaken to cover aspects such as completeness, absence of duplication, accuracy of spelling, correct choice of relationships, etc.

13.3.3 Construction of a multilingual thesaurus

Three approaches for constructing a multilingual thesaurus can be considered. The choice of approach is dependent on the available human, time and financial resources. It should also take into account the context of the intended application, for example whether monolingual thesauri already exist for the intended domain and whether they have already been used to index the documents to which access is required.

a) Translation of a monolingual thesaurus

The translation approach is a popular and relatively economical way of constructing a multilingual thesaurus. It is especially cost effective if the documents to be searched have already been indexed with the monolingual thesaurus. If this approach is adopted, however, the language of the original monolingual thesaurus acquires the status of dominant language. The translation process produces multiple target language versions, with terms and identical relationship structures that might not always correspond to the cultural and conceptual expectations of the user communities in the target languages.

b) Merging of several distinct monolingual thesauri

The merging of distinct monolingual thesauri is a complex process. It is more respectful of the integrity of the different conceptual and linguistic systems that must be reconciled, as it allows all languages to assume the role of source language in turn. However, as the various thesauri used as sources typically exhibit different levels of specificity and pre-coordination, this approach is the most challenging from a management perspective.

c) Simultaneous construction of the various language versions of the multilingual thesaurus

This third approach enables all languages in the thesaurus to have equal status. Each language becomes source and target language in turn, and can have an impact on the establishment of hierarchical and associative relationships. The project should be managed in such a way that the terminological and structural adaptations required to maintain cross-language equivalence are not always made in the same language. Careful coordination and teamwork should be applied from the start, to ensure a successful outcome.

13.3.4 Involving experts and stakeholders

The following benefits can be obtained from involving other people at various stages of the work. All should be pursued.

a) Specialist terminology can be difficult to grasp and have surprising nuances of meaning. Someone who knows a specialized area well can quickly check whether the terminology chosen is appropriate and relationships correctly appointed. If terms are applied in ways that are unacceptable to experts within the user community, the thesaurus will not easily gain acceptance. In principle, therefore, it is a good idea to involve experts in checking as work proceeds. However, terms typically have more than one meaning, and the meaning favoured by one expert might not meet the approval of another. Furthermore, as the concept most useful for information retrieval purposes might be slightly broader or narrower than the term preferred for representing it, some experts find it difficult to accept the perceived "inaccuracy". The experts chosen to assist should be well briefed in the purposes and conventions of a controlled vocabulary (see Clause 4). The thesaurus developer should be prepared to guide and mediate all contributions.

b) It is difficult for one person to make all the correct choices at the first attempt. A great deal can be gained from discussing difficult terms among an editorial team, with a degree of lateral thinking. If the team consists of one editor, it might be beneficial to set up a committee or panel to review sections of the work in progress, or to consult selected colleagues ad hoc. Optionally, the editorial review board might be asked to approve all elements of the thesaurus before they are issued.

c) The occasional involvement of senior management and/or other stakeholders can be an asset in securing continued funding, acceptance and implementation of the finished product.

d) Involvement of the users can greatly enhance buy-in. People with a stake in the project are much more likely to use it.

e) The involvement of experts in all the languages of a multilingual thesaurus, from the start, greatly improves the prospects for acceptability to communities speaking the various languages.

13.4 Introduction to the thesaurus

All thesauri should provide a comprehensive introduction that clearly states the purpose of the thesaurus and why, how, and by whom the thesaurus was developed. The following points should be addressed:

a) the subject field(s) covered, with core and fringe areas identified;

b) the languages in which the thesaurus is presented;

c) confirmation of which national and/or international standards have been followed;

d) the meaning of all conventions, abbreviations and any punctuation marks used in non-standard ways;

e) the total number of terms, with separate totals of preferred and non-preferred terms;

f) any rules governing the selection of concepts, including complex concepts;

g) the rules adopted in selecting the forms of preferred terms, including a reference to any style manual that is followed, and the rules for deciding their interrelationships;

h) the filing rules employed, following (and quoting) an appropriate national or international standard where possible;

i) a statement on the updating policy, including the frequency, dates and procedures, and the name and address of the responsible agency to whom comments and suggestions should be sent;

j) references to, and acknowledgements of, sources used in compiling and reviewing the thesaurus.

Whenever possible, these points should be illustrated by examples and accompanied by definitions of any technical terms used in the introduction.

If the thesaurus was designed for a particular system or application, the introduction should also explain how to use it in that context.

13.5 Dissemination

13.5.1 Integration with an electronic system

The needs of people using a thesaurus, either for indexing or for searching, should be considered. Access is most convenient if the thesaurus with all its terms, relationships and scope notes is fully integrated with the indexing, browsing and searching functions, as discussed in 16.3.

Some indexing and/or searching systems have an integrated thesaurus maintenance module. If this is used for the thesaurus development, no import or export is required, and no special preparations are needed. However, if the integrated system has no mechanism for exporting the thesaurus in a standard format, difficulties will occur if the need arises to change systems or to make the same thesaurus available for additional applications.

Whether or not routine dissemination is needed, a capability for exporting the thesaurus in a standard format should be available. Indexing and searching systems that use the thesaurus should be capable of importing the thesaurus in the same format. Guidance on formats is given in Clause 17 and Annex B.

13.5.2 Electronic publication

The thesaurus may also be made available in stand-alone mode, i.e. not linked to any particular application. An electronic thesaurus may be distributed on CD-ROM, on a website, an intranet or the Internet. The applicable browse functions are described in 16.3.

Even when a thesaurus is published on a network without a direct linkage to any particular application, some search applications on the same network (an intranet or the Internet) can use it for real-time interrogation while a search is in progress. To make this easier, the thesaurus should be made available in XML format, accessible as an XML "namespace".

Certain characters have restricted use in XML (e.g. < > & ' "). These characters should be avoided where possible as described in 6.4.2. Where, however, they are needed for user acceptability (e.g. Boyle's Law, women's rights) they should be encoded by the thesaurus management software as set out in References [57] and [58] when generating the XML output.

13.5.3 Limited distribution of hard copy

The thesaurus may be made available in hard copy for limited distribution under the following circumstances:

a) a full-scale print run would be too expensive;

b) most users prefer an electronic version;

c) the thesaurus is strictly an in-house product;

d) the thesaurus editorial team find a printout convenient for display and/or note-taking.

Realistically, this option is a simple extension of the routine practice of obtaining reports from the thesaurus database on demand. It offers the flexibility to tailor-make the report(s) for ad hoc needs, for example to include or exclude the "housekeeping fields", to issue a subset for a particular user group, to choose between the various styles of display, etc. Such reports may be distributed electronically, leaving individual users to print them out as required.

13.5.4 Conventional publication

While conventional publication has largely been superseded by electronic publication, it is still appropriate in some cases, if justified by user demand. Printed displays have some advantages over electronic ones, particularly in locations where access to a computer is limited. Furthermore, conventional publication can bring the thesaurus to the notice of a different, important audience.

When an organization has decided to develop and publish a new thesaurus, notification of intent should be announced in a professional journal in the relevant field(s), as well as through relevant electronic discussion forums.

13.5.5 Deposit with a clearinghouse

A copy of the first and any subsequent edition of a published thesaurus should be deposited with an appropriate national depository library. If the thesaurus is in English, or includes English as one of its languages, a copy should also be deposited with the North American clearinghouse in Toronto[4)].

4) Subject Analysis Systems Collection (SAS) is held at The Inforum, Faculty of Information, University of Toronto, 140 St George Street Toronto, Ontario M5S 3G6 Canada. See details at http://discover.library.utoronto.ca/general-information/libraries/FISINFORUM/. SAS records are included in the University's catalogue.

13.5.6 Website directories

Lists of thesauri are shown on a number of websites available through the Internet. Some of these sites publicize thesaurus projects in advance of publication. Additionally, the Dublin Core Metadata Initiative (DCMI) has encouraged several ongoing projects to establish metadata registries, including a "terminology registry" intended for controlled vocabularies among other resources[31]. On or before publication, the thesaurus developer should send full details of the thesaurus, whether electronic or printed, to the administrator of suitable websites or registry services. In the case of an electronic thesaurus available through the Internet, the service might provide a direct link to it.

13.6 Updating

13.6.1 General

Thesaurus maintenance should be active throughout the lifetime of the product. It is essential to provide for changes, both in the natural evolution of terminology and in the scope of the databases or other resources that the thesaurus serves. The need for updating begins the day a thesaurus is issued, if not before. Suggestions for change might be particularly frequent in the first few months after issue, as that is the time when the vocabulary is undergoing its first serious testing. Editorial procedures should be in place from the start.

While change is essential to cope with emerging requirements, it is important to recognize that every change potentially impairs retrieval performance for the "backfile" (i.e. for the resources already indexed with the original vocabulary). When a new term is introduced, it is available for searching the database from that date on, but relevant resources already in the system will not have used the term in their indexing. If the database management software allows automatic or semi-automatic correction of backfile records, changes to existing preferred terms can be implemented. However, other types of changes might be less susceptible to retrospective fixing. Therefore, a compromise might be required between meeting new needs and maintaining stability.

If the thesaurus is multilingual, the impact of the proposed change on all the languages should be considered, before approval.

13.6.2 Suggestion procedure

An easy mechanism for suggesting changes should be provided for all users of the thesaurus, in particular indexers. The mechanism provided may be a form, either printed or electronic. The form should invite people to suggest a new term or other change that is required, and provide as much background information as possible. Useful background information includes a definition of the concept required, any references to its occurrence in the literature, any synonyms or alternative forms of expression, any relationships with terms already present in the thesaurus, and a justification for why the change is needed.

Indexers often discover the need for a new term while processing documents for input to a database. As there might not be time to consult the thesaurus editor before completing the indexing, the proposed term should be applied as a "candidate preferred term". The candidate terms become available for searching, but are not required to pass thesaurus validation. The thesaurus editor should regularly review all the candidates proposed (see 13.6.3) and may admit them to the thesaurus management system with "candidate" status. As and when a candidate is approved, the appropriate adjustment should be made to the status of the term. Candidates not approved should be deleted or corrected as appropriate.

13.6.3 Review procedure

At regular intervals the thesaurus editor should review all the suggestions received and candidate terms proposed. Furthermore, the editor should monitor the usage frequency of preferred terms in the database(s) served. Any that are used too often or too little are candidates for deletion or amendment.

User queries may also be used as an aid to maintenance. Terms used in unsuccessful searches can be examined to see if they are valid concepts for the thesaurus. Heavy use or non-use of a term in queries can indicate that it should be considered for division or deletion, respectively.

Many of today's information retrieval systems offer a free-text search for finding information resources and keep a log of the transactions, including search terms entered by the users. Such transaction logs can serve as an objective guide to user preferences and habits, indicating, for example, which subject areas need more depth of coverage, and whether simple or more complex concepts are sought. Sometimes they suggest additional non-preferred terms for a concept already in the thesaurus, or even the replacement of an existing preferred term with a synonym that is more frequently sought. Any transaction logs that are available should be fully exploited as a guide to thesaurus maintenance.

The suggestions, user search terms and indexers' candidates should be researched thoroughly. Careful study can show that the suggested concept was already present in the thesaurus under another name. As the most common proposals are for new preferred terms, they rarely give consideration to all the alternative or complementary actions required, such as adding a lead-in entry, changing a scope note, adjusting hierarchical or associative relationships, etc. They might also add unacceptably to the level of specificity. The editor should flesh out each proposal to a fully coherent and consistent set of amendments before approval can be given. If the thesaurus is multilingual, the proposal should be worked out, with adaptations wherever necessary, in all the languages.

In a small establishment the editor may need to make all the decisions, although consultation with users is advisable, particularly if the thesaurus is used for several applications. A note of all the proposed changes can be circulated to users with an invitation to comment; alternatively, a review board, with representation from each major application, can be consulted. This board should meet regularly or organize a regular electronic forum. After fully researching and fleshing out the proposals, the editor should present them to the board. The implications for each application should be taken into account before a decision is reached.

13.6.4 The nature of changes

While the main focus of updates tends to be adding new terms, other types of change should not be neglected. Any of the following listed in a) to n) can occur:

a) adding a new concept, represented by a preferred term;

b) adding a new non-preferred term as a lead-in entry;

c) amending a preferred term or a non-preferred term;

d) deleting a preferred term;

e) making a preferred term a non-preferred term of another preferred term;

f) deleting a non-preferred term;

g) reversing the preference between a preferred and a non-preferred term;

h) adding or changing a scope note, definition, history note or editorial note;

i) merging two concepts into one;

j) splitting a concept into two or more, which can sometimes involve selecting an existing non-preferred term to be a preferred term;

k) altering the hierarchical structure;

l) moving a branch of the hierarchy from one place to another;

m) adding or removing associative relationships;

n) almost any combination of a) to m), sometimes involving different changes in the different languages of a multilingual thesaurus. For example, the addition of a non-preferred term in one language may or may not need to be reflected in other languages.

It is essential that the original editorial criteria continue to apply, with the added criterion of respecting the convenience of existing users. Deletions are of particular concern, especially if the term has been used in the past. Instead of deleting a preferred term, the term should be demoted to have non-preferred status and point to the appropriate preferred term, with a history note showing the date until which it was valid. The corresponding preferred term should also have a history note pointing out the change.

While the editor should craft an appropriate history note when terms are changed significantly, the date of input and date last modified should be filled automatically by the software. Similarly, the identifier for new terms should be added automatically.

The changes should be implemented at the same time in all the languages of a multilingual thesaurus.

13.6.5 Dissemination of updates

13.6.5.1 General

The mechanism by which updates are distributed should be planned before the thesaurus is first issued. Communication of changes can be complex, particularly when the attributes and relationships of one term have changed in several ways. While it is straightforward to send out a list of new terms, it is difficult if the users need to know where the terms fit into the hierarchical structure, particularly if several different types of display need to be distributed. Assuming that a simple list of new terms is acceptable to users, the likelihood is that the new terms will be ignored at the time of using the thesaurus because they have not been fully integrated into the look-up procedure.

The simplest way of disseminating changes is to distribute a new edition of the thesaurus. If this is carried out electronically, costs are kept to a minimum. Furthermore, it avoids the difficulty of communicating the nature of the changes and ensures that users all have access to the same vocabulary authority when selecting terms. Additionally, there is a need to communicate the details of the changes both to human users and to computer systems that will import the update.

13.6.5.2 Reader-friendly updates

For human users, notification may take the form of a list in alphabetical order, possibly complemented by one in classified order, or a categorized list showing all the new preferred terms that have been added or altered, new lead-in entries, etc. Feedback should be invited on the format. The list can be made more or less elaborate depending on the views expressed. Some user groups might want to see the full details of every new or amended term and relationship. Electronic displays can incorporate devices such as tracking or highlighting to draw attention to the changes. All users should be encouraged to note the most important changes, such as new terms and deletions, in their printed copy if they use it. For printed thesauri, this can be facilitated by producing them in loose-leaf form whereby amendments take the form of complete pages that can replace outdated ones.

13.6.5.3 Updates for computers

When a copy of the thesaurus is held in another computer system, the easiest option for updating the copy is often to delete the old one and import a whole new thesaurus with the changes implemented. However, this might not be acceptable if the target system links thesaurus terms with other data, such as postings to bibliographic records or mappings to another vocabulary. For computer systems such as these, reports or files of all the entities, attributes and relationships (see Clause 15) that have been deleted, added or modified should be provided as a minimum.

To support these requirements, the thesaurus management system should be capable of recording the date on which the entities, attributes or relationships were last modified, and using this date to extract a report of the modified data. Alternatively, an external log file can record all additions, deletions and changes to entities, attributes and relationships.

For importing systems, some of the changes will be insignificant, while others will be important. For this reason, some thesaurus management systems record the nature of the change as well as the change itself. This allows the importing computer system to separate relatively straightforward changes, e.g. spelling changes, from complex changes, e.g. splits of a term into two terms. As a standard categorization of the possible changes does not exist, the procedures for achieving this should be developed ad hoc.

13.6.5.4 Tracking changes

The importance of the date last modified field has already been noted in connection with updates for computers. It is sometimes also needed for reader-friendly updates, as is the date of input. The date should be entered in the standard format yyyy-mm-dd in accordance with ISO 8601. The concept identifier and/or term identifier is also vital, in case the spelling of a term is changed. (Without the identifier a computer has no way of finding out that a given term is intended to replace another with a different spelling.) All of the attributes mentioned should be filled in automatically to avoid human error.

It is highly recommended that a note be kept of all terms deleted. However, few thesaurus management packages have a mechanism for retaining a list of terms deleted and providing reports of them. Therefore, the editor should keep a note, recording the identifier as well as the term itself.

14 Guidelines for thesaurus management software

14.1 General

Thesaurus management software should be capable of managing, importing and exporting the data as described in Clause 15. It should be chosen on the basis of the requirements of the project in hand (see 13.1.5). While the characteristics and functions outlined in 14.2 to 14.10 are usually needed, exceptions can occur. Also, additional features are often required. For example, it can be advantageous if the software is capable of simultaneously handling more than one thesaurus.

NOTE No guidance is given in this document on the criteria applied to software selection in general, such as the availability of good documentation, training and other forms of support, the general user-friendliness of the interface and an acceptable price. Considerations applicable to vocabulary mapping will be covered in ISO 25964-2.

14.2 Size and character limitations

The software should not impose the following limitations.

a) There should be no limitations on the number of terms in the vocabulary that prevent it expanding to the size needed. The same is applicable to other elements such as node labels, scope notes, etc.

b) There should preferably be no limitations on the length of terms or of node labels, notes, etc. While relatively few terms exceed 40 characters, some need 100 characters or more.

c) There should be no limitation on the number of hierarchical levels admissible, nor on the number of hierarchical, associative or equivalence relationships possessed by any preferred term.

d) The software should be capable of handling all the characters in the Universal Character Set, as defined in ISO/IEC 10646, when used in any of the text elements (including terms, node labels and notes).

e) The editor should be able to choose upper or lower case characters as appropriate.

f) For multilingual thesauri, there should be no limit on the number of languages that can be included. It should be possible to apply all the desired languages to any or all of the text elements, such as terms, node labels and notes.

14.3 Relationships between terms and between concepts

The software should conform to the requirements specified in a) to i) below.

NOTE Where this section refers to the English tags BT, NT, RT, USE, UF, SN the corresponding tags in other languages should be applied in those languages (see Table 2 in Clause 3).

a) It should not permit the coexistence of duplicate terms for the same language. Any duplicates entered should be rejected upon entry or at least submitted to the editor for correction, amalgamation, the addition of a qualifier, or other remedial action. The matching algorithm for duplicate detection should be capable of customization so that, for example, typographical differences such as italics or capitalization may be ignored for the purposes of duplicate detection.

b) It should support the basic relationships BT/NT, RT/RT, USE/UF as defined in Clauses 8 and 10.

c) It should support the reciprocity requirements described in Clauses 8 and 10, i.e. if Concept A has a BT relationship with Concept B, then Concept B should have an NT relationship with Concept A and vice versa; if Concept C has an RT relationship with Concept D, then Concept D should have an RT relationship with Concept C; if Term E has a USE relationship with Term F, then Term F should have a UF relationship with Term E and vice versa. Preferably, the software should make the reciprocal relationships available automatically in response to the editor's insertion one way round, but as a minimum the software should issue a warning if any non-reciprocal relationship is present.

d) When a term or concept is amended or deleted, the change should be propagated automatically in all places in which that term or concept appears related to another (whether as BT, NT, RT, USE or UF). In the event of term or concept deletion, all relationships to and from that term or concept should be deleted. However, if deletion of a term or concept or relationship would leave any concept without at least one broader term/concept, or any non-preferred term without a preferred term, the editor should be warned.

e) There should be no limitation on the number of relationships a given concept or preferred term can have with others. Thus one concept can have any number of BTs, NTs and RTs, and one preferred term can have any number of non-preferred terms.

NOTE Certain combinations are inadmissible (see g) below).

f) It should be possible to set up user-defined reciprocal relationships, for example to distinguish between different types of BT/NT or different types of associative relationship.

g) Validation checks should prevent the entry of inadmissible relationship combinations, as follows:

- If two terms or concepts already have one of the basic relationships, no other basic relationship between the same terms or concepts is admissible.
- If Concept A has BT Concept B, none of the concepts in the BT hierarchy above Concept B should be admissible as BT, NT or RT of Concept A.
- Non-preferred terms (i.e. any term with a USE or USE+ relationship to another term) cannot have any BT, NT, RT or UF relationships.
- In the case of the USE+ relationship occurring, the software should check that the relationship is at least ternary.

h) No relationship from a term or a concept to itself is admissible, i.e. all the basic relationships are irreflexive.

i) Only one preferred term should be admissible for each concept, in each language of the thesaurus.

14.4 Notes applying to terms or concepts

The software should support notes as follows:

a) It should support entry of a scope note, associated with any concept.

NOTE Scope notes can be of any length.

b) If a note (of any type) makes reference to another term or concept in the vocabulary, the software should preferably support addition of a marker or hyperlink to the record for that term or concept. The software should check the validity of the link target.

c) It should support the setting up of user-defined notes associated with any term or concept, for example history notes, editorial notes etc.

14.5 Codes and notation

The software should support subject categories, codes and notation as follows:

a) It should be possible to associate at least one code, number or other type of notation with any term or concept, concept group or array (see 12.2.2 and 15.2.18). Preferably more than one coding/numbering type should be supported.

b) It should be possible to associate one or more subject category/ies with any term or concept, concept group or array (see 12.2.2 and 15.2.18).

c) It should be possible to assign a unique identifier to each term and to each concept. Preferably, the assignment of the identifier(s) should be automatic whenever a new term and/or concept is entered, in such a way as never to duplicate any of the existing identifiers or identifiers of terms or concepts previously deleted. The identifier should not change when term or concept attributes or relationships are modified in any way.

d) It should be possible to output the vocabulary using the sequence of each type of notation, coding or identifier.

14.6 Node labels

The software should support node labels as follows:

a) Node labels are not regarded as thesaurus terms or concepts (see Clause 11, especially the examples in Figure 4) and, therefore, are not subject to the relationship requirements under 14.3. Furthermore, they do not need to be unique and so should not be subject to the duplicate control described in 14.3 a).

b) The software should have means for locating a node label in displays in the correct position relative to any parent term and to the highest level terms that come within the corresponding facet or array.

14.7 Status of languages

The software should allow each language of a multilingual thesaurus to have equal status, avoiding predominance of one of the languages over the others. For example:

a) The number of non-preferred terms applying to any one concept in each language should not be determined by the number present in another language, and the non-preferred terms in different languages should not be required to correspond to each other.

b) The existence of a scope note for a given concept in one language should not require a corresponding scope note in any of the other languages.

14.8 Data import/export

It is essential to be able to export and import data so that the thesaurus content is not restricted to the software or application initially used to develop it. Import/export formats are discussed in Clause 17. The following mechanisms should be provided.

a) Bulk import of datasets from existing vocabularies, comprising terms, scope notes, node labels, standard relationships between the terms and concepts, and all other attributes of the terms, concepts and node labels.

   — All of the mentioned features should be retained after import, as well as all characters from the Universal Character Set wherever used.

b) Producing reports or exporting the vocabulary, including all terms, scope notes, notation and standard relationships between the terms and concepts.

   — It should also be possible to export editor-defined subsets, such as non-preferred terms only, or preferred terms with their scope notes and NTs only, or preferred terms of concepts that have no hierarchical or associative relationships, etc.

c) Batch edit/delete facility.

   — It should be possible to delete or edit batches of records in the same way, preferably using a facility native to the software. If this is not possible, an option might be to use functions of the underlying database management system, or to export the selected records, edit them externally and re-import. This technique can be used, for example, to add the same history note or relationship to a batch of terms, or perhaps a mapping to terms in another vocabulary. If this method is used, the customary validations should be applied on import.

d) Exporting all terms that have been amended after a given date, with or without their full details.

   — The option should be available of selecting only certain types of amendment, for example only new terms, or only those in a particular language, or of including all terms that have undergone any change of attributes or relationships. It should also be possible to report all terms that have been deleted from a certain date.

e) Outputting thesaurus displays, either in hard copy or on the screen.

   — It should be possible to choose between a variety of sequences and layouts for the display, of which the alphabetical display is essential and either a hierarchical or a classified display is highly recommended. See Clause 12 for a description of the various types of display.

f) When sibling terms in alphabetical, hierarchical or classified displays are presented in alphabetical order, either the word-by-word or the letter-by-letter convention may be used. The same convention should then be applied throughout the thesaurus. (An explanation of the conventions can be found in ISO 999; see also the example below.) However, it should be possible to force a different sequence when appropriate (see Clause 11).

   EXAMPLE

Word by word sequence	Letter by letter sequence
national insurance	national insurance
national parks	nationality
nationality	national parks

14.9 Editorial navigation and support

The software should support the following editorial functions.

a) The thesaurus editor should be able to locate the full details of any term (and the corresponding concept) by direct entry of that term, preferably only needing to enter a portion of the term, with the software assisting the selection by identifying all terms that contain the portion.

b) The editor should be able to locate the full details for any term (and the corresponding concept) by browsing via the term/concept relationships. The browse facilities should preferably enable him/her to start off from a list of top terms, but in the absence of this he/she should at least be able to navigate from any term to the record for any of its NTs, BTs, RTs, USEs or UFs.

c) The editorial interface should allow viewing of the full hierarchical context of a term or concept, at the same time as the term itself, with all its notes, codes and relationships presented for editing.

d) It should be possible to move a concept (together with all its NTs, NTs of those NTs, etc.) easily from one hierarchical position to another, preferably using drag-and-drop. This requires breaking one set of BT/NT relationships and making another. Any RT relationships thus invalidated should be drawn to the editor's attention.

e) Editing capabilities should include standard word processing facilities, such as the ability to add, modify and delete characters without rekeying an entire field. A spell-checking function might also be useful.

f) When an editor takes steps to delete a term or concept, the software should first seek confirmation that the deletion of that term or concept is intended, before completing the action.

g) When a relationship is to be established between two terms or concepts already present in the vocabulary, it should be possible to achieve this by navigation and selection rather than by having to key in the whole of a term already known to the system. This is important to prevent errors, as well as to support efficiency.

h) The editor should be able to switch easily from a display in one language to the corresponding display for any of the other languages in the thesaurus.

14.10 Editorial safeguards

The software should embody the following safeguards.

a) Editorial changes should be made in the first instance to a master database from which outputs are derived periodically for downstream processes such as indexing or resource discovery applications.

b) If more than one person is editing the master database simultaneously, an inbuilt mechanism should prevent simultaneous write access to the same records.

c) Security/password controls should prevent editorial changes from unauthorized persons.

d) Preferably, the software should allow for different levels of access so that provisional changes can be made, which are not finally admitted until approved.

e) For each editor, a roll-back (undo) function should allow progressive reversal of the most recent editorial changes he/she has entered.

f) A log should be kept so that the database can be restored from an earlier version.

14.11 Housekeeping tools

It should be possible to obtain reports of the numbers of terms with particular characteristics, particularly the total number of preferred and non-preferred terms.

As described in 13.6.3, the frequency of use of thesaurus terms in indexing and searching should be monitored. Also, the thesaurus management software should be capable of importing and storing the usage data if another tool does not already perform this function. One or more user-defined fields [see 14.4 c)] or custom attributes (see 15.2.16) might be needed for storage of the data.

15 Data model

15.1 General

This clause models the underlying structure of the data that can be conveyed in a thesaurus conforming to earlier clauses of this part of ISO 25964. For multilingual thesauri, however, this data model applies only to those that are fully symmetrical. If data structures differ between languages it might be necessary to employ mapping techniques, at least for the parts of the thesaurus where this occurs. Mapping techniques will be discussed in ISO 25964-2.

The XML schema described in Annex B is derived from this model and may be used when sending or receiving thesaurus data in electronic form. The model represents a logical structure that should be maintained, although it does not necessarily represent the way in which data is held within a computer system.

The basic recommendations in Clauses 5 to 12 are quite simple, needing only a simple model. But there are also a number of optional features that require much greater complexity in the model. For thesaurus exchange to be successful in a wide range of situations, it is important to have a format hospitable to sophisticated as well as simple thesauri.

The model presented here accommodates the full range of options described in Clauses 5 to 12. Some elements of this model are shown as optional and need not be used when exchanging thesauri that lack the optional features. Software systems importing data may ignore optional elements that they do not support, and assume appropriate default values (which might be null) for optional elements that they do support but that are not present in the imported data, while interpreting the remaining elements consistently and correctly.

The model is presented first in Figure 15 using UML (Unified Modelling Language[42][43][44]) conventions, but stops short of specifying the thesaurus behaviours and class methods that might be needed in a software implementation specification. UML is highly expressive and discourages ambiguity, but as some users might be unfamiliar with its conventions, an alternative tabular presentation is added at 15.3. The data elements and attributes in the UML diagram and the data tables are identical in substance, though some details of relationships are not made explicit in the tabular presentation. In case of doubt, the diagram should be regarded as the authority.

UML defines its own special conventions and terminology for modelling purposes. In the present clause, the term "class" is used in the UML sense to describe a set of objects that have similar structure, behaviour and relationships. Similarly, UML gives special meanings to the expressions "String", "Boolean" and "aggregation". More information about UML can be found from references in the Bibliography. In Reference [44], Chapter 5 is a guide to UML notation, and Section 4.5.4 describes detailed semantics.

The notes on the model given in 15.2 apply to the tabular presentation in 15.3, as well as to the diagram in Figure 15.

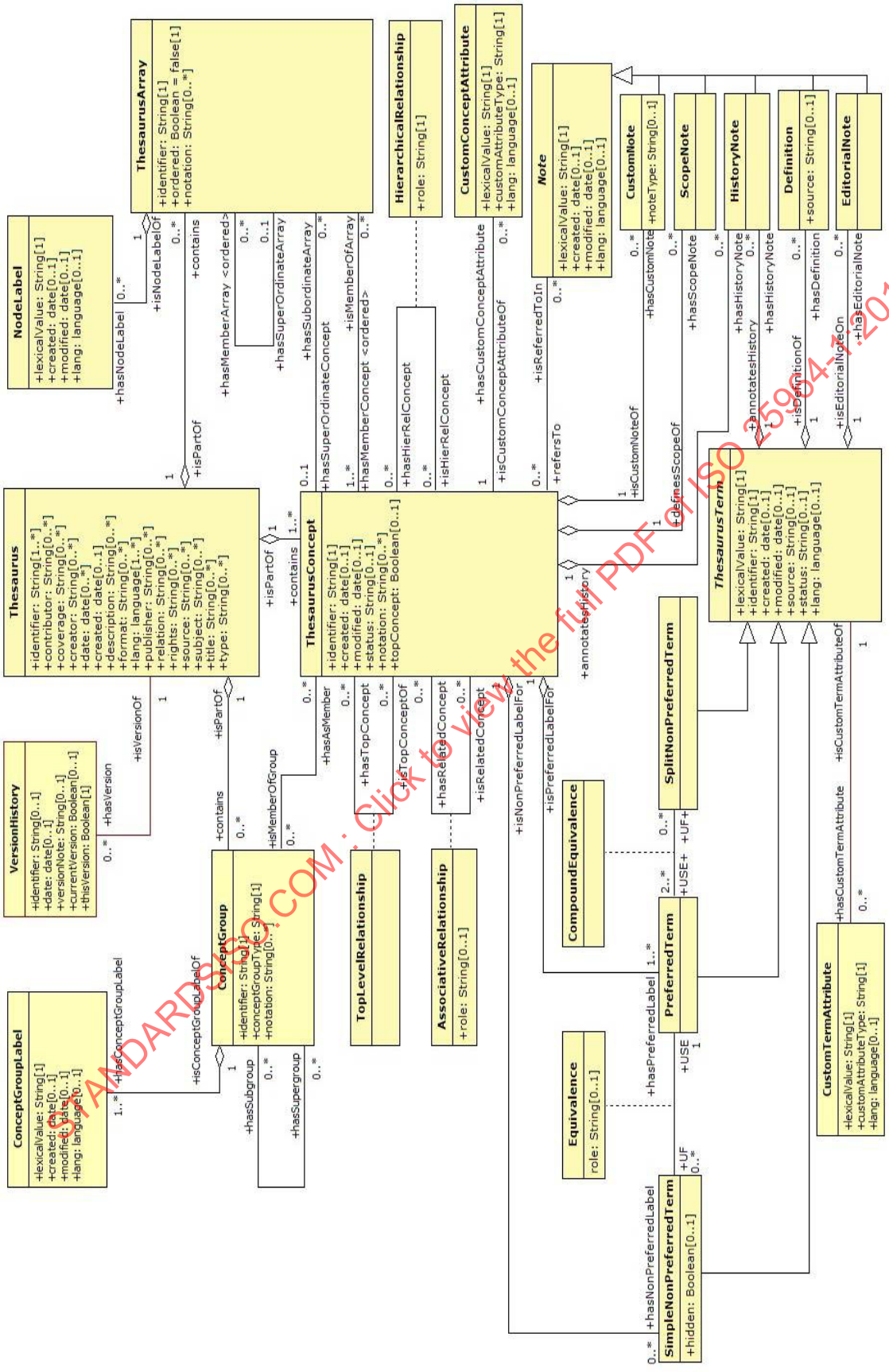


Figure 15 — UML model

15.2 Notes on the model

15.2.1 Naming conventions

The names of all attributes, including some Dublin Core metadata elements (see Reference [33]) that appear in the model and in the derived XML schema, follow the Dublin Core "lowerCamelCase" convention (i.e. no spaces between words and each word except the first starting with a capital letter). Names of UML classes use "UpperCamelCase" (i.e. no spaces between words and each word starting with a capital letter).

15.2.2 Symbols for associations

The diamond symbol in the diagram represents an "aggregation", i.e. a "has-a" relationship. This symbol is used to indicate that in the derived XML schema the definition of an object will be nested in its parent object.

The triangle symbol indicates "generalization", i.e. an "is-a" relationship, linking a general class with more specific types of that class, which inherit all the attributes of the parent class.

Normal associations (without a diamond) indicate a relationship of the type key/keyref in XML, where each instance of one class contains a reference key that links it to an instance of another class. Note that the XML schema described in Annex B does not actually enforce the key/keyref constraints. The reason for this relaxation of the constraints is that enforcement would have the effect of invalidating XML output files that contain only an extracted portion of the thesaurus rather than a whole thesaurus.

15.2.3 Concept

Each concept in the thesaurus is represented by one preferred term per language, and by any number of non-preferred terms. The notation, scope note and broader/narrower/related term relationships apply to the concept as a whole, rather than to its preferred term. A unique identifier may be assigned to each concept. In some systems, the concept is identified only by its preferred term or by the identifier of its preferred term, but this has disadvantages if the spelling of the term changes.

15.2.4 Identifier

Thesaurus, ThesaurusConcept, ThesaurusTerm, ThesaurusArray and *ConceptGroup* each have a unique identifier that is used in the XML schema described in Annex B. The identifier and date attributes of *ThesaurusTerm* are essential for the provision of a good updating service because if the spelling of a term changes, a constant Term identifier facilitates continuity during successive updates. The use of a concept identifier is strongly recommended to promote interoperability among networked search applications.

15.2.5 Dates

The date created and date modified are strongly recommended but formally optional attributes of several classes. They should be given in YYYY-MM-DD format, in line with ISO 8601. In the case of *Thesaurus*, another option is to use the simple attribute "date", in which case the value will be interpreted as applying to the date when the version to which it is attached was issued. The same format should be used.

15.2.6 Versions and version history

The class VersionHistory optionally allows any copy of a thesaurus to carry a record of versions or editions that have been created.

Although the class is optional and might not be needed when only one version exists, adoption is highly recommended as soon as there is more than one. Each version should be identified by an identifier or a date or both.

versionNote can be used to explain the nature of the version, e.g. whether it is an updated version, an extract or a translation, or to explain its relationship to other versions.

currentVersion is a Boolean (true/false) flag to indicate for each version whether it is still current or whether it has been superseded or withdrawn. More than one version can be current simultaneously.

thisVersion is a Boolean flag to indicate which of the versions listed is the one to which this history is attached.

Distinguishing between those thesaurus variants which are "versions" and those which should be identified as separate works is not always easy. In the present context, versioning control relates mainly to successive updating of an original thesaurus. Thus:

a) Updates done or adopted by the original owner/creator, i.e. new concepts, terms, relationships, notes, changes to existing parts of the thesaurus and special editions for different purposes, should normally be treated as new versions of a given thesaurus.

b) True subsets, even though they might have a subtitle such as "Abridged version", are more conveniently treated as concept groups (see 15.2.18), since they are based on the content of an existing version.

c) Similarly, different language versions, provided they correspond to the original work and share the same conceptual structure (even though they might be referred to as "English version", French version", etc.), should be treated as belonging to the same thesaurus version. For data exchange purposes, they should be distinguished by using the language attribute of *Thesaurus*.

d) Any work incorporating changes of terms, concepts or relationships, new translations, simplification/reductions of relationships and similar, if it is issued by others than the owners/original creators, should be identified as a separate thesaurus. Changes in the Thesaurus attributes creator, created, publisher, rights or title might indicate that it is a separate work. In such cases, it should be given a distinct identifier and linked to the original work by data in attributes such as relation, description and source.

15.2.7 Status

This is an optional attribute of *ThesaurusConcept* and *ThesaurusTerm,* which records whether they are, for example, approved, candidates, superseded or deprecated (see 13.6.2).

15.2.8 Notation

This is an optional attribute of *ThesaurusConcept, NodeLabel* and *ConceptGroup* (see 12.1.3 and 12.2.5.2).

If the thesaurus uses an expressive notation, then applying it to node labels will allow them to be shown in the correct place in hierarchical displays.

If there is no expressive notation, some other means must be found of outputting the node labels correctly in the display, such as a sort code attribute that is not displayed to users. In this event, the same attribute will be required at the display level for preferred terms.

The system of notation used for *ConceptGroup* may be quite distinct from that used for *ThesaurusConcept*, and one of these systems may be present without the other.

15.2.9 Source

This is an optional attribute of *Thesaurus* and *ThesaurusTerm*, which can be used to note the reference work or individual who contributed the term in question.

15.2.10 Language codes

Language is an optional attribute of *ThesaurusTerm, Note* and *NodeLabel* and a mandatory attribute of *Thesaurus*. It should be given as an alpha-2 code selected from ISO 639-1 if present in that list, or an alpha-3 code from ISO 639-2 if not. These codes may be extended where necessary with the additional codes described in RFC 4646[45] and listed in the IANA subtag registry[35] (see 12.4.5).

15.2.11 Dublin Core elements

The *Thesaurus* class has a number of attributes intended for assigning Dublin Core metadata[33] to a file using the ISO 25964 format. As well as identifier, date, source and language already described, these include contributor, coverage, creator, description, format, publisher, relation, rights, subject, title and type. All except language and identifier are optional.

15.2.12 Multiplicity of Preferred Term

The model allows more than one preferred term per concept, to provide for multilingual thesauri where one is permitted per language. For monolingual thesauri the multiplicity is reduced to one.

15.2.13 Top level concepts

Clause 12 makes brief mention of "Top Terms", i.e. terms representing top concepts, as an optional feature of alphabetical displays and single term displays (see 12.2.2 and 12.1.3), linking a concept to the top concept of the hierarchy in which it occurs. The class *TopLevelRelationship* enables this.

A browse capability that starts by presenting editors with a list of top terms is also recommended in 14.9 b). The attribute **topConcept** addresses this, specifying that a concept is at the top of a hierarchy, i.e. it has no broader concepts. This makes it easier, on importing a thesaurus, to pick out all the top level concepts from which to build a navigational tree.

15.2.14 Notes

Zero or more notes of each type in each language may be recorded.

Editorial notes are useful for entries such as "Review this term after the company merger complete" or "This term is mentioned in the scope note of Concept X" (see 5.3) or "Check spelling with expert A". Notes such as this, and several of the attributes, are more useful for housekeeping than for user consultation. A choice can be made whether or not to allow users to view them.

15.2.15 Note references

The association between *Note* and *ThesaurusConcept* enables any note for one concept to refer to any other concept in the thesaurus. This capability is particularly useful for scope notes (see 5.3).

15.2.16 Custom attributes

The model includes classes *CustomConceptAttribute* and *CustomTermAttribute* for custom attributes of concepts and terms. These enable recording of custom information about concepts and terms.

These are included as separate classes rather than as normal attributes so that the administrator of the thesaurus management system can specify the values of custom attributes that can be assigned. The classes have an attribute customAttributeType, allowing the administrator to specify which type of attribute is being used. Values of customAttributeType should normally be taken from a controlled list.

15.2.17 Arrays and node labels

Although all thesauri have arrays of sibling terms, it is not always necessary to model them explicitly and so this class is optional. The *ThesaurusArray* and *NodeLabel* classes of the model are needed particularly to support the generation of displays incorporating node labels and/or non-alphabetic sequences of sibling terms (see examples in Figures 4, 6, 7, 8 and 12; see also 12.4.3).

The attribute **ordered** enables non-alphabetic sequences when necessary.

In each language of the thesaurus a maximum of one node label per array should be present.

Some systems organize the display by treating node labels as though they were concepts, with relationships to broader and narrower concepts. This is done only for display purposes and not for the normal functions of broader and narrower relationships. Only a concept can have true broader and narrower relationships.

15.2.18 Concept groups

Many thesauri group concepts using a classification structure that exists in parallel to the hierarchies of thesaurus concepts based on BT/NT relationships. Groups created by the classification are often based on disciplines, subject areas or areas of business activity. They are sometimes called "subject categories", "themes", "domains", "groups", "subsets" or "microthesauri". The model provides for all of these by providing the classes *ConceptGroup* and *ConceptGroupLabel* and the specific type may be indicated by the attribute *conceptGroupType*. In general, there is not a BT/NT relationship between a *ConceptGroup* and the concepts that it contains.

Concepts may be gathered into *ConceptGroups* from many different facets or hierarchies of the thesaurus, and the notation used for the classification into groups may be quite distinct from any notation that might be used for the concepts themselves. Groups may have subgroups, being nested to any level. Each group should be given one verbal label per language.

If it is important to exclude selected non-preferred terms from a given subset of concepts, it might be necessary to mark the selected terms, which may be done using a custom attribute (see 15.2.16).

15.2.19 Treatment of facets

Facets are provided for in the data model although not modelled explicitly, since different thesauri apply facet analysis in different ways. If facet names are included as top terms, they should be treated as normal preferred terms applying to top level concepts (see 15.2.13). Alternatively, they may be the names of concept groups (15.2.18), and/or they may appear in node labels (15.2.17).

15.2.20 Relationship roles

The classes *Equivalence, HierarchicalRelationship* and *AssociativeRelationship* each have a "role" attribute. This can be used to convey the common relationships USE/UF, BT/NT and RT/RT. It can also be used for the optional relationship types described in 10.2.1, as well as custom relationship types as suggested in 10.4. Each occurrence of a relationship is expressed in one direction only, between instances of a "source" class and a "target" class. For example, "Concept A has a broader concept, namely Concept B", where the role is "broader". If the complementary relationship "Concept B has a narrower concept, namely Concept A" needs to be expressed, then this is shown as a separate occurrence with the role "narrower".

For custom relationship types, the text given in the "role" attribute should be composed of (a) the name of the parent relationship type, followed by (b) the symbol forward slash "/", and finally (c) the name of the custom relationship type. If necessary, custom relationship types can be subdivided further in the same way.

EXAMPLE

The text in the 'role' attribute of HierarchicalRelationship may be one of the following, where NTX indicates some further subdivision of NTI:

NT
NT/NTP
NT/NTI
NT/NTG
NT/NTI/NTX
BT
BT/BTP
BT/BTI
BT/BTG
BT/BTI/BTX
etc.

15.2.21 Split non-preferred terms

The classes *SplitNonPreferredTerm* and *CompoundEquivalence* enable representation of complex concepts by a combination of terms, as described in 8.5.

15.2.22 Role of relationship tags

Figure 15 shows a relationship between *SimpleNonPreferredTerm* and *PreferredTerm,* with the tags USE and UF, which are described in Clauses 3 and 8 in the context of conventions and displays for human readers. In the context of data exchange, it could be argued that this relationship is redundant, since all the preferred and non-preferred terms for a given concept are linked to that concept. However, some redundancy is retained to facilitate exchange of data for simpler thesauri and to make it easier to exchange update information and other specialized subsets of the data.

Another detail concerns "USE…+". In displays for human readers, there is always a gap between "USE" and "+", to accommodate the term in between (e.g. "Coal mining USE coal + mining"). But if the tag is to be used behind the scenes by computers, it needs to be compressed, in the style "coal mining USE+ coal"; “coal mining USE+ mining". For this reason, the gap is omitted in the diagram.

15.3 Tabular presentation

The elements of the model shown in Figure 15 are listed in Tables 4 to 9, based on the five classes *Thesaurus, ThesaurusArray, ThesaurusConcept, ThesaurusTerm* and *Note*. Other subsidiary classes have been shown within these tables. The brief text in the description columns is complemented by the more extensive notes in 15.2. See also definitions for many of the terms in Clause 2.

The Type column indicates the data type, using the UML data types "String" and "Boolean" where applicable, the XML data types "Date" and "Language" where no UML type is applicable, and "Class" in the case of an initial or associated class in the model.

The columns headed "Occurs" describe whether the elements are mandatory and/or repeatable as shown in Table 3.

Table 3 — Definitions of Occurs as used in Tables 4 to 9

Occurs	Meaning
1	mandatory, not repeatable (one occurrence only)
1..*	mandatory, repeatable (one or more occurrences)
0..1	optional, not repeatable (zero or one occurrence)
0..*	optional, repeatable (zero or more occurrences)

Table 4 — Attributes and associations of *Thesaurus*

Element	Type	Occurs	Description
Thesaurus	Class	1	The thesaurus as a whole
Attributes of *Thesaurus*			
identifier	String	1..*	Identifier for the thesaurus as a whole
contributor	String	0..*	Person or organization who contributed to the thesaurus
coverage	String	0..*	Spatial or temporal coverage of the thesaurus
creator	String	0..*	Person or organization primarily responsible for making the resource
date	Date	0..*	Any date associated with the thesaurus
created	Date	0..1	The date the thesaurus was created
modified	Date	0..*	A date when this version was modified
description	String	0..*	An account of the thesaurus
format	String	0..*	The file format or physical medium of the thesaurus
language	Language	1..*	Codes showing languages supported by the thesaurus
publisher	String	0..*	Entity responsible for publication
relation	String	0..*	A related publication
rights	String	0..*	Copyright or other rights information
source	String	0..*	Resource from which the thesaurus was derived
subject	String	0..*	Index terms indicating the subject content
title	String	0..*	Title of the thesaurus
type	String	0..*	The genre of the vocabulary, e.g. "thesaurus"
Associated classes of *Thesaurus*			
ThesaurusArray	Class	0..*	An array of sibling concepts forming part of the thesaurus
ThesaurusConcept	Class	1..*	A concept forming part of the thesaurus
ConceptGroup	Class	0..*	A group of concepts selected by some specified criterion, such as relevance to a particular subject area
VersionHistory	Class	0..*	May provide confirmation of currency as well as relationships with other versions of the same thesaurus

Table 5 — Attributes and associations of *ThesaurusArray*

Element	Type	Occurs	Description
ThesaurusArray	Class	0..*	An array of sibling concepts forming part of the thesaurus
Attributes of ThesaurusArray			
identifier	String	1	A unique identifier for the array
ordered	Boolean	1	A true/false label indicating whether the order of concepts in the array has to be maintained. Default is "false".
notation	String	0..1	A symbolic notation attached to the array, which may be used for sorting and display
Associated classes of ThesaurusArray			
hasSuperOrdinateArray	Class	0..1	A higher-level array of which this array is a member
hasMemberArray	Class	0..*	A lower-level array that is a member of this array *An array needs to have at least one member, but this can be a concept or another array.*
hasSuperOrdinateConcept	Class	0..1	A higher-level concept to which this array is subordinated
hasMemberConcept	Class	0..*	A concept that is a member of this array *An array needs to have at least one member, but this can be a concept or another array.*
NodeLabel	Class	0..*	A label showing the basis on which the concepts in the array have been chosen and ordered
— Attributes of NodeLabel			
— lexicalValue	String	1	The wording of the node label
— created	Date	0..1	The date when the node label was created
— modified	Date	0..1	The date when the node label was last modified
— lang	Language	0..1	A code showing the language of the node label. This should be included if the thesaurus supports more than one language

Table 6 — Attributes and associations of *ThesaurusConcept*

Element	Type	Occurs	Description
ThesaurusConcept	Class	1..*	A concept that is a member of the thesaurus
Attributes of ThesaurusConcept			
identifier	String	1	A unique identifier for the concept
created	Date	0..1	The date when the concept was created
modified	Date	0..1	The date when the concept was last modified
status	String	0..1	An indication of whether the concept is candidate, superseded, etc.
notation	String	0..*	A symbolic notation attached to the concept, which may be used for sorting and display
topConcept	Boolean	0..1	A true/false label indicating whether the concept is at the top of a hierarchy, i.e. has no broader concepts
Associated classes of ThesaurusConcept			
PreferredTerm	Class	1..*	The term used as a label for this concept. There should be one preferred term per language.
SimpleNonPreferredTerm	Class	0..*	Alternative terms by which this concept could be sought
ScopeNote	Class	0..*	A note defining or clarifying the scope of the concept within this thesaurus
HistoryNote	Class	0..*	A note recording changes to this concept within this thesaurus
CustomNote	Class	0..*	A note of any other kind relating to this concept
— ***Attributes of CustomNote***			
— noteType	String	0..1	Specification of a particular kind of note
Note	Class	0..*	A note of any kind that contains a reference to this concept although it applies more directly to another concept
isMemberOfArray	Class	0..*	An array of which this concept is a member
hasSubordinateArray	Class	0..*	An array that is subordinate to this concept
CustomConceptAttribute	class	0..*	An additional attribute of a concept
— ***Attributes of CustomConceptAttribute***			
— lexicalValue	String	1	The wording of the custom attribute
— customAttributeType	String	1	Specification of particular kinds of attribute
— lang	Language	0..1	The language of the LexicalValue
TopLevelRelationship	association class		
— ***Classes linked by TopLevelRelationship***			
— hasTopConcept	Class	0..*	The concept at the top of a hierarchy containing this concept
— isTopConceptOf	Class	0..*	A concept in the hierarchy of which this concept is top
HierarchicalRelationship	association class		
— ***Attributes of HierarchicalRelationship***			
— role	String	1	Specification of the kind of hierarchical relationship that the target concept has to the source concept, e.g. BT, BTI, NT, NTP etc.

Element	Type	Occurs	Description
— ***Classes linked by HierarchicalRelationship***			
— hasHierRelConcept	Class	0..*	Target concept, linked to this source concept by the relationship specified in "role", e.g. if role is "BT", this is a link to a broader concept of this concept
— isHierRelConcept	Class	0..*	Source concept to which this target concept is linked by the relationship specified in role, e.g. if role is "BT", this is a link to the concept of which this is a broader concept
AssociativeRelationship	association class		
— ***Attributes of AssociativeRelationship***			
— role	String	0..1	Specification of the kind of associative relationship that the target concept has to the source concept. This will usually be "RT", but relationship could be asymmetric, e.g. "CAUSE" as discussed in 10.4.
— ***Classes linked by AssociativeRelationship***			
— hasRelatedConcept	Class	0..*	Target concept, linked to this source concept by the relationship specified in role, e.g. if role is "CAUSE", this is a link to a concept that is a cause of this concept
— isRelatedConcept	Class	0..*	Source concept to which this target concept is linked by the relationship specified in role, e.g. if role is "CAUSE", this is a link to the concept of which this concept is a cause

Table 7 — Attributes and associations of *ConceptGroup*

Element	Type	Occurs	Description
ConceptGroup	Class	0..*	A group of concepts selected by some specified criterion, such as relevance to a particular subject area
Attributes of ConceptGroup			
identifier	String	1	A unique identifier for the group
conceptGroupType	String	1	A label identifying the type of group, e.g. "microthesaurus", "theme", or "subject category"
notation	String	0..1	A symbolic notation attached to the group, which may be used for sorting and display
Associated classes of ConceptGroup			
ThesaurusConcept	Class	0..*	A concept that is a member of this group
hasSuperGroup	Class	0..*	A higher-level group of which this group is a member
hasSubGroup	Class	0..*	A lower-level group that is a member of this group
ConceptGroupLabel	Class	1..*	A label providing a verbal description of the group. A group should have one label per language.
— ***Attributes of ConceptGroupLabel***			
— lexicalValue	String	1	The wording of the label
— created	Date	0..1	The date when the label was created
— modified	Date	0..1	The date when the label was last modified
— lang	Language	0..1	A code showing the language of the label. This should be included if the thesaurus supports more than one language

Table 8 — Attributes and associations of *ThesaurusTerm*

Element	Type	Occurs	Description
ThesaurusTerm	Class	1	A term in the thesaurus by which a concept could be sought
Attributes of ThesaurusTerm			
lexicalValue	String	1	The wording of the term
identifier	String	1	A unique identifier for the term
created	Date	0..1	The date when the term was created
modified	Date	0..1	The date when the term was last modified
source	String	0..1	The person(s) or document(s) from which the term was taken
status	String	0..1	Indication of whether the term is candidate, approved, etc.
lang	Language	0..1	A code showing the language of the term. This should be included if the thesaurus supports more than one language
Associated classes of ThesaurusTerm			
HistoryNote	Class	0..*	A note recording changes to this term within this thesaurus
EditorialNote	Class	0..*	A note for use by the thesaurus editors during the editing process
Definition	Class	0..*	A note giving definitions of a term, not necessarily limited to the scope of the concept labelled by the term in this thesaurus
— Attributes of Definition			
— source	String	0..1	The person(s) or document(s) from which the definition was taken
CustomTermAttribute	Class	0..*	An additional attribute of a term
— Attributes of CustomAttribute			
— lexicalValue	String	1	The wording of the custom attribute
— CustomAttributeType	String	1	Specification of particular kinds of attribute
— lang	Language	0..1	The language of the lexicalValue
Specializations of ThesaurusTerm			
PreferredTerm	Class		A term that has been chosen to label a concept
SimpleNonPreferredTerm	Class		A term that serves as an alternative label for a concept, but that is not the preferred term
— Attributes of SimpleNonPreferredTerm			
— hidden	Boolean	0..1	A yes/no flag to show whether the term may be excluded from some forms of output, e.g. for misspellings of a term.
SplitNonPreferredTerm	Class		A term that represents a combination of two or more thesaurus concepts
— Associations of PreferredTerm and SimpleNonPreferredTerm			
— Equivalence	association class		
— Attributes of Equivalence			
— role	Class	0..1	Specification of a kind of equivalence relationship. This will normally be USE, linking the source SimpleNonPreferredTerm to the target PreferredTerm
— Classes linked by Equivalence			
— PreferredTerm	Class	1	Target term, linked to this source term by the relationship specified in role, e.g. if the role is "USE", this is a link to a PreferredTerm that is to be used in place of the current SimpleNonPreferredTerm
— SimpleNonPreferredTerm	Class	0..*	Source term to which this target term is linked by the relationship specified in role, e.g. if the role is "USE", this is a link to the SimpleNonPreferredTerm for which the current PreferredTerm should be used instead
— Associations of PreferredTerm and SplitNonPreferredTerm			
— CompoundEquivalence	association class		

Element	Type	Occurs	Description
— ***Classes linked by CompoundEquivalence***			
— SplitNonPreferredTerm	Class	0..*	A non-preferred term naming a complex concept that is represented by a combination of preferred terms. See 8.5.
— PreferredTerm	Class	2..*	One of the two or more PreferredTerms that together represent the CompoundNonPreferredTerm

Table 9 — Attributes and associations of *Note*

Element	Type	Occurs	Description
Note	Class	0..*	A piece of text giving additional information about the term or concept to which it is linked
Attributes of Note			
lexicalValue	String	1	The wording of the note
created	Date	0..1	The date when the note was created
modified	Date	0..1	The date when the note was last modified
lang	Language	0..1	A code showing the language of the note. This should be included if the thesaurus supports more than one language
Associated classes of Note			
ThesaurusConcept	Class	0..*	A concept referred to in the note, other than the one to which it is directly attached
Specializations of Note, for particular types			
ScopeNote	Class	0..*	A note defining or clarifying the scope of the concept within this thesaurus
HistoryNote	Class	0..*	A note recording changes to a concept or a term within this thesaurus
Definition	Class	0..*	A note giving a definition of a term, not necessarily limited to the scope of the concept labelled by the term in this thesaurus
EditorialNote	Class	0..*	A note for use by the thesaurus editors during the editing process
CustomNote	Class	0..*	A note of any other kind relating to a concept
— ***Attributes of CustomNote***			
— noteType	String	0..1	Specification of a particular kind of note

16 Integration of thesauri with applications

16.1 Introduction

The prime function of a thesaurus is to support information retrieval. The thesaurus is typically invoked at two main stages in the retrieval chain: at the time of indexing and at the time of search. While software suites exist that manage thesaurus development and maintenance, indexing and search in one integrated system, it is more common to manage these as separate applications. Thesaurus data therefore need to be exchanged smoothly between these applications, avoiding errors and inconsistencies as updating continues.

The scenario has become even more complex now that widespread electronic networks and the almost universal use of computers in information management present opportunities for sharing data and information on a scale not previously imaginable. From a single computer connected to the internet, one can simultaneously search different collections of books, articles, photographs, data sets, or museum artefacts residing on remote computers all over the world. To enable thesaurus exploitation in this sort of scenario, the

systems involved need interoperability, i.e. the capability of multiple systems or components to exchange and use information.

This clause describes:

a) the features and functions that enable thesaurus interoperability across systems and networks, and

b) the features and functions needed within a single indexing or search application that uses a thesaurus.

It does not describe the features and functions needed for applications that use two or more linked thesauri; such applications will be covered in ISO 25964-2.

16.2 Interoperability needs for thesauri

As described in 14.8, it should be possible to export and import a thesaurus, in part or in whole, from/to the application where it is maintained. A standard format should be used for the import/export (see Clause 17).

When a thesaurus has been updated, it should be possible to export all the records that are new or have changed in some way, including any that have been deleted.

When a thesaurus update is implemented, the system should be able to incorporate the new or changed terms without losing any data assigned to the previous terms, where still valid. For example, there may be links from the existing terms to resources indexed with them.

To make the thesaurus available to users of a network such as the Internet, it should be published on that network, using a standard format (see Clause 17).

A published thesaurus should be updated regularly. This might require import and implementation of an updating file from the application where the master version is maintained.

When the thesaurus has been published on a network, a standard protocol (see Clause 18) should be used to extract all the data, or a subset or some selection from it. For example, a search application might need to select some or all of the information associated with a particular concept or group of concepts. A unique identifier should be associated with each concept to facilitate distributed search applications.

For the convenience of users and wider dissemination, the thesaurus should be registered with any vocabulary registries serving the anticipated user community.

16.3 Integration with indexing and searching applications

16.3.1 If a thesaurus is to be used for information retrieval, it should be capable of integration with the systems in which indexing and searching take place. For a printed thesaurus, integration is achieved by the process of a person looking through the pages, selecting terms and re-entering them into the information retrieval system with the correct syntax. With an electronic thesaurus, these functions should be supported in a user-friendly way.

Unless thesaurus management is already integrated in one system with the indexing and search applications, the first requirement is to export the thesaurus data from the thesaurus management system into the information retrieval system. Formats and protocols for this purpose are discussed in Clauses 17 and 18.

16.3.2 All systems that are involved in the handling of thesaurus data should be able to support all of the characters likely to be present in the data. These typically include parentheses in term qualifiers, and sometimes other punctuation marks. In the case of a multilingual thesaurus, all of the characters in the script and/or alphabet of each of the languages need to be supported.

16.3.3 For indexing and searching applications, except when these are fully automated without provision for user interaction, facilities for browsing and searching the thesaurus are fundamental. These should include the following.

a) The ability to search for terms containing any word, word combination or character string.

b) The ability to search for terms and concepts using any notation and/or identifiers present.

c) The ability to choose or switch the display language of a multilingual thesaurus, or limit the search and display to a single specified language.

d) The ability to display the equivalents (preferred or non-preferred) for a given term, including cross-language equivalents if the thesaurus is multilingual.

e) The ability to move from the record for one concept to that of any of its broader, narrower or related concepts, generally achieved by setting up hypertext links. This facility might remove the need for a notation to link the alphabetical and classified or hierarchical displays.

f) The ability to expand and contract levels when viewing a classified or hierarchical display. To gain an initial overview of the structure it should be possible to start with a view of the top level, then expand progressively and selectively downwards. It should also be possible to expand or contract fully at one stroke.

g) The ability to switch from one style of display to another, for example, from the classified to the alphabetical display. Preferably it should be possible to view them simultaneously.

16.3.4 When integrated with a searchable document collection and when provision is made for the user to refer to the thesaurus, the displays should also show the number of postings against each preferred term. While browsing the thesaurus, a user should be able to move directly to the searching phase. The following facilities should be provided.

a) The ability to select one or more terms from a thesaurus display to be used in a search.

b) The ability to build a search statement by copying terms from the thesaurus without retyping (for example by double-click or drag-and-drop). Help with applying valid search syntax (e.g. selecting Boolean operators) should ideally be provided.

c) The ability to select subtrees from the thesaurus to be used in a search, i.e. to search for a given term and all its narrower terms combined with Boolean OR, sometimes called the "explode" function.

d) Selection of a non-preferred term should lead to substitution with the corresponding preferred term in the search statement (ensuring that the searcher is notified of what has been done). In the event that the user enters a non-preferred term into a search statement without first browsing the thesaurus, he/she should be offered the option of converting it to the corresponding preferred term.

e) During the process of search refinement, the numbers of postings in any of the thesaurus displays should be adjusted to show the numbers that apply in the subset of the collection already selected.

16.3.5 It is possible also to exploit a thesaurus within a search engine without exposing it for user browsing. Facilities 16.3.4 (c and d) may be adapted for this type of use.

EXAMPLES

a) When a user enters a term similar to one or more thesaurus terms, the search engine might respond "Did you mean….?" thus offering the user the nearest matches found in the thesaurus.

b) When a user enters a term similar to several thesaurus terms, the search engine might respond by retrieving results for all of these terms, and clustering them term by term. (For example, if a thesaurus has two terms "turkey (meat)" and "Turkey (country)", and the user searches simply for "Turkey", the search results may be presented in two groups, corresponding to each of the thesaurus terms respectively.)

c) The user might be given the option to "broaden" or "narrow" the search without being told exactly how this is done.

d) In the event of retrieving very few results, the search might be automatically expanded to broader, narrower and related terms.

When a search term is automatically converted, substituted, translated, replaced by a broader or narrower term, etc., the user should be informed that this has taken place.

16.3.6 The equivalence relationships in a thesaurus may also be exploited to search a database that has not been indexed with the thesaurus. For example, if a thesaurus includes "motor-bikes", "motor-cycles" and "motorcycles" among the preferred and non-preferred terms for the same concept, then a search for any of these terms can be expanded to include the others. See also the discussion of synonym rings in ISO 25964-2.

16.3.7 When integrated with an indexing or cataloguing system, the following functions should be provided.

a) The ability to copy one or more terms directly into an indexing form or template without retyping.

b) Validation of terms entered directly, against the thesaurus. In the event that a non-preferred term is entered, this needs to be converted automatically to the corresponding preferred term with a notification to the indexer. When a string is entered that corresponds to the start of one or more thesaurus terms, the indexer should be offered all the matching preferred terms from which to choose.

c) A facility to submit candidate terms to the thesaurus editor (see 13.6).

16.3.8 If the system is for automatic indexing, rather than for human indexers to use, a means of linking each preferred term with the algorithms or rules that drive its selection should also be provided.

17 Exchange formats

Interoperability between different computer applications requires use of a common format. Four existing formats are particularly well known:

— **MARC (MAchine-Readable Cataloguing)**[51] – This set of standards, based on ISO 2709, is widely used for exchange of bibliographic information between libraries. While the prime application of the standards is to exchange cataloguing records, the set provides also for "authority records", i.e. those drawn from thesauri or other types of controlled vocabulary. Of particular relevance are:

- *MARC 21 Format for Authority Data* (available in full and concise versions)[52];
- *MARC 21 Format for Classification Data* (available in full and concise versions)[53];
- MADS (Metadata Authority Description Schema) a MARC21-compatible XML format for the type of data carried in records in the MARC Authorities format[54].

MARC formats expressed in XML may also follow ISO 25577.

— **SKOS** – The SKOS (Simple Knowledge Organization Systems) format[60][61] is a W3C recommendation particularly applicable to Semantic Web developments, and is encoded using XML and RDF (Resource Description Framework). As well as thesauri, it is being developed for classification schemes, taxonomies and other vocabulary types.

— **Zthes** – Zthes[69] was originally developed as an application profile of Z39.50 to enable exchange of thesaurus data, but has since evolved, and has a version that can be used with SRU (Search and Retrieval via URL; see Clause 18).

— **DD 8723-5** – This BSI Draft for Development provides an XML schema that may be used for the exchange of a thesaurus compliant with BS 8723[25]. The schema is available at http://schemas.bs8723.org/.

In addition to the four mentioned, others are sometimes used to meet the needs of specific domains.

Also relevant is ISO 16642, the Terminological Markup Framework (TMF) for computer applications in terminology. This framework is designed to provide guidance on the basic principles for representing data recorded in terminological data collections. It includes a meta-model and methods for describing specific terminological markup languages (TMLs) expressed in XML. It is designed to support the development and use of computer applications for terminological data and the exchange of such data between different applications. Data categories and methods for the specification of data structures are specified in ISO 12620 and other related International Standards.

As none of the existing formats is adequate to convey all the features of a thesaurus using the model in Clause 15, the XML format described in Annex B is recommended.

It seems likely that several different standard formats will coexist for some time. One way to minimize the inconvenience of multiple formats is to convert the data by a suitable transformation, even though some information is liable to be lost if a thesaurus is transformed into a format that does not support all the features of the source format. Some transformations are currently under development. The Schemas website (http://www.niso.org/schemas/iso25964) provides a space on which future developments can be found.

18 Protocols

18.1 General

A protocol is an essential tool in the communication process enabling thesauri to be used across digital networks. Several different protocols are already in existence, enabling the live querying of thesauri for purposes of indexing and/or retrieval as described in 16.3. New developments are frequent in this fast-moving field and so a comprehensive listing would be inappropriate. As each protocol has strengths in a different set of circumstances or application environment, those listed in 18.4 and 18.5 may be considered for a particular application. The selection and use of the protocol should be based on the needs of the application, especially the intended purpose and the software environment.

18.2 Purposes and use cases

It is important to distinguish between discovering or identifying a complete thesaurus and retrieving metadata about it, versus retrieving specific concepts and terms (or groups of concepts and terms). Protocols can be employed to present and apply thesauri, their member concepts, terms and relationships, to describe the meaning of terms and facilitate semantic interoperability. This is done for broader purposes of searching, browsing, discovery, translation, mapping, semantic reasoning, subject indexing and classification, harvesting, alerting etc. (see Reference [49]).

18.3 Application environment and architecture

Some existing protocols were designed specifically for thesauri (see 18.4) while others are general-purpose protocols that are being used (or adapted) to work with thesauri (see 18.5). The former type enables more powerful exploitation of thesaurus structure and relationships, while the general-purpose tools may facilitate integration with other functionality or existing skill-sets. It might also be possible to apply a thesaurus interface or profile to a general purpose protocol. In some cases, the limited set of function calls provided by a specific thesaurus protocol can offer advantages in hiding details of the underlying architecture or representation. On the other hand, sometimes the greater flexibility of a general purpose protocol can be exploited by developers familiar with it.

For a particular application it might be necessary to take into account the implementation platform, such as SOAP[62] or a lighter weight HTTP protocol such as REST[34] (and REST-like) XML-RPC[68] or JSON-RPC[38]. A few thesauri have recently been published as Linked Data[40], a connection mechanism reliant on use of RDF and URIs, which could be considered a variant of REST.

18.4 Thesaurus-specific protocols

Three protocols deserve mention:

— **SWAD-E SKOS API** – This is a Web Service API (Application Programming Interface), designed to provide access to thesauri and other simple knowledge organization systems (SKOS) via the Web[67]. It defines a core set of operations for programmatically accessing and querying a thesaurus. While intended for web service calls, the API itself remains independent of such concrete implementation details. The API could also be capable of adaptation to formats other than SKOS. Whereas the SKOS API was a product of the now complete Semantic Web Advanced Development (SWAD) Europe project[46], SKOS itself is maintained by the W3C Semantic Web Deployment Working Group (SWDWG).

— **Other APIs** - Various adaptations of the SKOS API are currently in use on a variety of platforms. There are also some similar but unrelated thesaurus protocols, both SOAP and REST based. Non-exhaustive lists are maintained currently at References [39] and [47].

— **ADL Thesaurus Protocol** – The Alexandria Digital Library (ADL) Protocol[36] is a protocol for accessing monolingual thesauri, adapted to use over the World Wide Web. It supports live querying of, and navigation around, the thesaurus. It offers a lightweight HTTP option for thesaurus access. The protocol's model of a thesaurus closely follows ANSI/NISO Z39.19[22] and the definition is specified in an XML schema.

18.5 General-purpose web database protocols used with thesauri

For some applications, one of the following general-purpose protocols could be appropriate:

— **Search Web Services (OASIS)** – These protocols are based on an underlying Abstract Protocol Definition (APD)[32], a reference model by which these and other protocols can be described. The most notable are OpenSearch[28] and Search and Retrieval via URL (SRU)[46] with its search language Contextual Query Language (CQL)[29]. Designed for use across unstructured documents, OpenSearch is simpler while SRU gives more control over query and results. SRU is an XML-focused search protocol for Internet search queries. SRU/CQL is currently being revised by an OASIS Search Web Services Technical Committee[41]. An earlier protocol known as SRW is now considered a variation of SRU ("SRU via HTTP SOAP"); SRU is REST-based whereas SRW is SOAP-based.

— **SPARQL** (Simple Protocol and RDF Query Language)[63] – An official World Wide Web Consortium (W3C) recommendation as of 2008, it has been developed and promoted by the W3C as a key semantic web technology. This RDF query language is used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. SPARQL contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions. It can be considered a general-purpose query language (as SQL is for relational databases) but can be used to query thesauri that have been represented in RDF, allowing integration with other RDF datasets.

— **Z39.50** – This protocol is more formally known as ANSI/NISO Z39.50, *Information retrieval (Z39.50): Application service definition and protocol specification*[23]. SRU/SRW can be considered off-shoots of the earlier Z39.50 protocol.

Annex A
(informative)

Examples of displays found in published thesauri

A.1 Introduction

To complement the Figures in Clause 12, this annex shows some alternative styles of display found in a selection of well-known thesauri published in print or online. Some of these have optional features not described in this part of ISO 25964 and some depart from the conventions recommended by this International Standard, for example by introducing non-standard tags. Nonetheless, all demonstrate ideas for how to display the data, which can be adopted in particular contexts.

A.2 Inspec Thesaurus

The printed version of the Inspec Thesaurus has two sections, showing two main forms of display respectively: alphabetical, as described in 12.2.3 and illustrated in Figure A.1, and hierarchical, as described in 12.2.4 and illustrated in Figure A.2. Table A.1 explains the tags used in the alphabetical section. Unlike the example illustrated in Figure 6 of 12.2.4, the *Inspec Thesaurus* does not group terms into facets at the top level and so the hierarchical display has a large number of top terms, of which just six appear in Figure A.2. As illustrated in Figure A.1, all the main entries for preferred terms in the alphabetical section include one or more lines for top term (TT), which helps the reader find the relevant place(s) in the hierarchical section. For example, the alphabetical display for “microphones” shows it has “acoustic equipment” as one of three top terms. Correspondingly, the hierarchical display for “acoustic equipment” includes “microphones” (marked in Figure A.2 using arrows) in a much more complete hierarchical context than can be shown in the alphabetical display.

Table A.1 — Tags used in Inspec Thesaurus alphabetical display

Tag	Meaning
USE	Use (see Table 1, Clause 4)
UF	Used for (see Table 1, Clause 4)
BT	Broader term (see Table 1, Clause 4)
TT	Top term (see Table 1, Clause 4)
NT	Narrower term (see Table 1, Clause 4)
RT	Related term (see Table 1, Clause 4)
CC	Classification code, referring to a separate Inspec Classification scheme
DI	Date of input, i.e. the date when the term in question was entered into the thesaurus
PT	Prior term, used to extend a search to a period earlier than the date of input of the current term, when necessary.