SLOVENSKI STANDARD
SIST ISO 25964-1:2013

01-julij-2013

Informatika in dokumentacija - Tezavri in interoperabilnost z drugimi slovarji - 1. del: Tezavri za pridobivanje informacij

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

Ta slovenski standard je istoveten z: **ISO 25964-1:2011**

ICS:

01.140.20	Informacijske vede	Information sciences

SIST ISO 25964-1:2013 **en,fr,de**

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.