SLOVENSKI STANDARD
SIST ISO 30042:2019

01-oktober-2019

Upravljanje terminoloških virov - TermBase eXchange (TBX)

Management of terminology resources -- TermBase eXchange (TBX)

Systèmes de gestion de la terminologie, de la connaissance et du contenu -- TermBase eXchange (TBX)

Ta slovenski standard je istoveten z: **ISO 30042:2019**

ICS:

01.020	Terminologija (načela in koordinacija)	Terminology (principles and coordination)
35.240.30	Uporabniške rešitve IT v informatiki, dokumentiranju in založništvu	IT applications in information, documentation and publishing

SIST ISO 30042:2019 **en,fr,de**

INTERNATIONAL STANDARD

ISO 30042

Second edition
2019-04

Management of terminology resources — TermBase eXchange (TBX)

Gestion des ressources terminologiques — TermBase eXchange (TBX)

Reference number
ISO 30042:2019(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: 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.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

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. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 37, *Language and terminology*, Subcommittee SC 3, *Management of terminology resources*.

This second edition cancels and replaces the first edition (ISO 30042:2008), which has been technically revised.

The main changes compared to the previous edition are as follows:

— industry-defined dialects consisting of data category selections corresponding to the needs of specific communities have been introduced;

— the XCS formalism has been removed and replaced with the requirement that the dialect be described and its name be declared on the root element of every TBX document instance;

— the DTD for the core structure has been replaced with a schema language-neutral definition;

— this document, containing the essential core and normative content, has been separated from ancillary content produced and distributed publicly by stakeholders;

— a simplified DCT (Data Category as Tag) style has been added alongside the traditional TBX style of DCA (Data Category as Attribute);

— xml namespaces have been introduced as a means for declaring the data categories used in a given TBX dialect (for DCT style).

NOTE Additional details about these and other changes are available on the TBX Info website[15].

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

This document defines a framework for representing structured terminological data, referred to as TermBase eXchange (TBX). Within this framework, a variety of industry standards, known as dialects, for specific types of terminology interchange scenarios and terminological data collections can be defined.

TBX is designed to support various types of processes involving terminological data, including analysis, descriptive representation, dissemination, and exchange in various computer environments. The primary purpose of TBX is the exchange of terminological data. For example, it facilitates:

— integrating or converting terminological data from multiple sources;

— comparing the contents of various terminological data collections;

— querying multiple terminological data collections by passing data through a common intermediate format on a batch or dynamic basis;

— placing data on an online site for download or public feedback;

— making terminology available dynamically in networked applications through a web service.

A TBX-compliant dialect can facilitate the exchange of terminological data between users, which include people such as translators and writers, as well as applications and systems, such as computer assisted translation tools and controlled authoring software. Therefore, it can be used for both human-oriented and machine-oriented terminological data processing. In this manner, it can enable the flow of terminological information between technologies and systems throughout the information production cycle, both inside an organization and with outside service providers.

TBX document instances of the same defined TBX dialect are interoperable and exchangeable with minimal loss or minimal need for negotiation, because they:

— adhere to the core structure;

— use, or have access to, the same data categories; and

— comply with the same dialect-specific constraints as other instances of the same dialect.

TBX document instances developed according to ISO 30042:2008 can be converted to comply with the current version of TBX by identifying a dialect with which the document instance complies and implementing the other changes in accordance with this document. A converter is available on the TBX Info website for such purposes[15].

NOTE Supplemental resources are available to assist implementers and users of TBX dialects on the TBX Info website[15].

TBX is limited in its ability to represent presentational markup (such as bold or italics). However, presentational markup can be autogenerated from descriptive markup in a TBX document instance.

Management of terminology resources — TermBase eXchange (TBX)

1 Scope

This document explains fundamental concepts and describes the metamodel, data categories, and XML styles: DCA (Data Category as Attribute) and DCT (Data Category as Tag). It also specifies the methodology for defining TBX dialects.

The audience for this document is anyone wishing to create a new dialect compliant with TBX. This document can also be used to analyze and to understand a terminological data collection or to design a new terminology database that complies with international standards and best practices. Typical users are programmers, software developers, terminologists, analysts, and other language professionals. Intended application areas include translation and authoring.

The TBX-Core dialect is described in detail in this document. All other industry-supported dialects are out of the scope of this document.

NOTE TBX dialects are defined by industry stakeholders. Any materials needed to implement currently shared dialects are publicly available as self-contained industry specifications (see for instance the TBX Info website[15]).

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 8601-1, *Date and time — Representations for information interchange — Part 1: Basic rules*

ISO 8601-2, *Date and time — Representations for information interchange — Part 2: Extensions*

ISO 12620, *Management of terminology resources — Data category specifications*

ISO 16642, *Computer applications in terminology — Terminological markup framework*

ISO 21720, *XLIFF (XML Localisation interchange file format)*

ISO/IEC 10646, *Information technology — Universal Coded Character Set (UCS)*

3 Terms and definitions

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

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at http://www.iso.org/obp

— IEC Electropedia: available at: http://www.electropedia.org

3.1
attribute class
group of one or more related attributes

3.2
child element
element that is subordinate to another element

3.3
classification element
element used to group data categories according to their function in a *concept entry* (3.5)

EXAMPLE <admin>, which expresses data categories for administrative information, such as <admin type="originatingPerson">.

Note 1 to entry: Data categories are instantiated as the value of the *type* attribute used with a given classification element.

3.4
complementary information
CI
information supplementary to that described in *concept entries* (3.5) and shared across the *terminological data collection* (3.29)

3.5
concept entry
terminological entry
entry
part of a *terminological data collection* (3.29) which contains the terminological data related to one concept

[SOURCE: ISO 16642:2017, 3.22, modified — new terms "concept entry" and "entry" added, synonym "TE" deleted, preferred term now is "concept entry" instead of "terminological entry", Note 1 to entry deleted.]

3.6
core structure
common structure and *data categories* (3.8) that are used in all TBX *dialects* (3.12)

Note 1 to entry: The core structure is compliant with ISO 16642 (TMF).

3.7
core structure module
core module
TBX-Core module
data category module (3.9) that contains only those *data categories* (3.8) that are part of the *core structure* (3.6)

3.8
data category
class of data items that are closely related from a formal or semantic point of view

EXAMPLE /part of speech/, /subject field/, /definition/.

Note 1 to entry: A data category can be viewed as a generalization of the notion of a field in a database.

Note 2 to entry: In running text, such as in this document, data category names are enclosed in forward slashes (e.g. /part of speech/).

3.9
data category module
module
list of permissible *data categories* (3.8) and constraints on them that are used in the design of a TBX-compliant *terminological data collection* (3.29)

EXAMPLE The TBX-Core module, which includes the data categories and structure common to all TBX dialects, the TBX-Min module, which adds a minimum number of data categories needed for simple glossaries, and the TBX-Basic module, which provides for a richer set of data categories.

3.10
DCA
data category as attribute
style of representing TBX data whereby most *data categories* (3.8) are expressed as the value of a *type* attribute on an XML element declared in the corresponding schema

EXAMPLE <termNote type="partOfSpeech">adjective</termNote>.

3.11
DCT
data category as tag
style of representing TBX data whereby most *data categories* (3.8) are expressed as XML generic identifiers

EXAMPLE <partOfSpeech>adjective</partOfSpeech>.

3.12
dialect
XML markup language that validates according to the *core structure* (3.6) of TBX and allows exactly those *data categories* (3.8) at those levels specified by a particular *data category module* (3.9) or set of data category modules and complies with all other relevant constraints

Note 1 to entry: "All other relevant constraints" refers to constraints that are necessary for the dialect in question but that are not expressible in either the core structure or the data category modules, such as date formats or conditional constraints. An example of this occurs in the dialect TBX-Basic, which requires a /definition/ OR a

/context/.

3.13
display name
name of a *data category* (3.8) as it appears on a software user interface or other medium

3.14
document instance
file containing *concept entries* (3.5) represented in a TBX *dialect* (3.12)

3.15
exchange
interchange
transaction involving exporting data from one *termbase* (3.28) and importing it into another termbase

3.16
global information
GI
technical and administrative information applying to the entire *terminological data collection* (3.29)

3.17
grouping element
XML element whose purpose is to group together a set of *child elements* (3.2)

3.18
object language
language being described

3.19
PID
persistent identifier
unique identifier that ensures permanent access for a digital object by providing access to it independently of its physical location or current ownership

[SOURCE: ISO 24619:2011, 3.2.4, modified – "persistent identifier" made second preferred term, Note 1 to entry deleted.]

3.20
private dialect
dialect (3.12) intended for private use that has not been described on a publicly accessible website

3.21
public dialect
dialect (3.12) that has been described on a publicly accessible website

Note 1 to entry: An example of a publicly accessible website is TBX Info[15].

3.22
root element
first element in a TBX *document instance* (3.14)

Note 1 to entry: The root element is <tbx>.

3.23
TBX agent
program or utility which generates, reads, edits, writes, processes, stores, renders or otherwise manipulates TBX-compliant *document instances* (3.14)

3.24
TBX export
process of creating a TBX *dialect* (3.12) *document instance* (3.14) from a *termbase* (3.28) or its subset

3.25
TBX import
process of inserting terminological data from one TBX *document instance* (3.14) into an existing *termbase* (3.29)

Note 1 to entry: The existing termbase can be empty or can already contain terminological entries.

3.26
TBX Module Description
TBXMD
formalism for identifying a set of *data categories* (3.8) and their constraints for a specific *data category module* (3.9)

3.27
term component
one of the words of a multi-word term, or one of the components of a single-word term (such as a morpheme)

3.28
termbase
terminology database
database comprising a *terminological data collection* (3.29)

3.29
terminological data collection
TDC
resource consisting of *concept entries* (3.5) with associated metadata and documentary information

EXAMPLE A TBX document instance, ISO 1087.

[SOURCE: ISO 16642:2017, 3.21, modified – in the definition, "concept entries" used instead of "terminological entries", Example added.]

3.30
working language
metalanguage used in *concept entries* (3.5) to describe *object language* (3.18) content

4 Fundamental principles and concepts

TBX refers to a framework consisting of two interacting components: a core structure and a formalism for defining data category modules. The core structure is expressed in a schema definition language such as RelaxNG (RNG). (The core is also represented by its own data category module.) This component-based approach supports the varying types of terminological data, or data categories, that are included in different terminological data collections. The approach mirrors the terminological markup framework (TMF) in that the core structure shall reflect the abstract data model of TMF in accordance with ISO 16642. In addition, it facilitates an explicit description of what any two dialects within the TBX framework have in common (the core structure) and how they differ (expressed in their respective data category modules). The combination of these two components defines a particular dialect. "TBX" without a dialect indicator is not a file format, it is not a terminology markup language, and it is not itself a dialect.

The TBX framework assumes that, because terminological data collections vary significantly, no one dialect would satisfy all user requirements. All dialects within the TBX framework adhere to the core structure, which is described in Clause 8. A RelaxNG schema for the core structure is referenced in Annex C, and the elements and attributes are described in Annex A.

Dialects can differ with respect to which data categories are allowed, and at what levels of a concept entry these data categories may occur. These constraints on the core structure are formally represented in one or more data category modules.

A data category module, or simply *module*, is a list of permissible data categories and constraints on them that are used in the design of a TBX-compliant dialect. Constraints are the permissible content of a data category (including subsets of a standard picklist value domain) and the levels of the concept entry where the data category may occur (see Clause 8).

NOTE Sample data category modules are available on the TBX Info website[15].

It is recommended that implementers of TBX adhere to ISO standards and industry guidelines governing the principles and methodologies of terminology management and the content and quality of terminological data collections, such as those described in Clause 2 and the Bibliography.

The information represented in a TBX document instance should be concept-oriented. The terms in a single entry are assumed to be synonymous unless otherwise noted.

Furthermore, if two systems both fully support a given TBX dialect, then information in that dialect can be preserved when terminological data is exported from one and imported into the other. In the context of TBX, interoperability implies this preservation of data. When different dialects of TBX are used by two systems, interoperability is reduced, and loss of data categories and their content can occur. Thus, claiming compliance to TBX without indicating the dialect does not guarantee any degree of interoperability.

5 Data categories

5.1 General

Data categories represent information about terms and concepts, for instance, /part of speech/ and /definition/. A list of data categories commonly used in termbases is provided in Annex B.
A description of these and other data categories is available in the data category repository DatCatInfo[10]. If another data category repository is used to describe data categories, it shall also comply with ISO 12620.

In running text, such as in this document, data category names are enclosed in forward slashes (e.g. /part of speech/). In a TBX document instance, and in the data category modules where data categories for a TBX dialect are declared, camel case (e.g. partOfSpeech) shall be used. Industry-accepted names for data categories in camel case are available in DatCatInfo. If the data categories in Annex B are used in a TBX document instance, the names in Annex B shall be used.

5.2 Data categories specified in the core structure module

In TBX, the following data categories are declared in the TBX-Core structure, and therefore are available to all TBX dialects, and are represented in the same way in all styles (see Clause 6):

— /date/

— /term/

— /note/

5.3 Data categories specified in the data category modules

All data categories not included in TBX-Core that are required for a particular TBX dialect are documented in the dialect's data category module or modules (see Clause 9). Such additional data categories may include, for example, /definition/, /part of speech/, /context/, /term type/, and so forth. A simple data category module such as the Min module introduces /definition/, but an additional module, such as Basic, then further extends the model by adding /context/ and other data categories. Hence, the TBX-Basic dialect consists of modules for TBX-Core, TBX-Min, and TBX-Basic.

6 DCA and DCT styles

There are two XML styles that may be used to represent terminological data: DCA (data category as attribute) and DCT (data category as tag). DCA is the style used for the examples in this document.

— DCA: <termNote type="partOfSpeech">adjective</termNote>

— DCT: <partOfSpeech>adjective</partOfSpeech>

In DCA style, most data categories are expressed as the value of the type attribute (in the above case: /part of speech/) of one of the elements declared in the core structure (in this case, <termNote>).

In DCT style, most data categories are reflected in the element generic identifier name. The corresponding core-structure element with which this data category is associated may optionally be indicated as the value of the *metaType* attribute. For example:

<partOfSpeech metaType="termNote">adjective</partOfSpeech>

In both cases, the value of the data category is the content of the XML element. These two styles are isomorphic. That is, they can be converted back and forth by an algorithm without loss of information. Even if the *metaType* attribute is omitted (e.g. <partOfSpeech>adjective</partOfSpeech>), the two representations can still be converted from one to the other if the algorithm has access to a table that indicates the core-structure element associated with each data category.

DCA style emphasizes the similarity among TBX dialects. DCA also allows all TBX dialects to be validated, at a first level, against the same schema (the core structure), by using a general-purpose XML parser.

DCT style looks more familiar to XML users who are accustomed to distinct element names rather than refinement of elements through attribute values.

NOTE Additional information about DCT style is available on the TBX Info website[15].

7 Dialects

7.1 General

Few terminology collections or applications use exactly the same set of data categories. TBX is a flexible framework because it allows user groups to select their own data categories. By doing so, they can create their own dialect adapted to their requirements. A TBX dialect complies with the core structure and implements one or more defined data category modules.

7.2 Dialect naming

Dialect names shall start with the "TBX-" prefix and end with a dialect indicator, such as "Basic", i.e., TBX-Basic. Although dialects are not standardized, industry groups and companies have in the past declared their own data models for purposes of sharing in public environments. Public dialect names can be published and thus made available for collaborative use, for instance on the TBX Info website[15].

All TBX dialects are built upon the TBX-Core module, which contains the essential data categories described in Clause 8. TBX dialects extend the core by adding a set of data category modules. For instance, one common industry dialect, TBX-Basic, consists of three modules: TBX-Core, TBX-Min and TBX-Basic (see 9.6).

If a dialect has been extended through the addition of one or more data category modules, a meaningful name qualifier shall be added to the dialect name separated by a period ".".

EXAMPLE TBX-Basic.Seo, where Seo is the name of a module which adds one or more data categories to those in the TBX-Basic dialect, in this case to include data categories for search engine optimization (SEO).

NOTE 1 Suggested subset relationships between and among public TBX dialects or private extensions thereof are available on the TBX Info website[15].

NOTE 2 Modules and dialects both begin with the "TBX-" prefix and are distinguished by using the descriptors "module" or "dialect" in conjunction with their name.

7.3 Example of a dialect

This subclause describes a fictitious TBX dialect called TBX-Sample dialect. For illustrative purposes, this dialect allows minimal terminological information.

The TBX-Sample dialect is defined as the combination of the TBX-Core module plus the TBX-Fiction module:

TBX-Sample dialect = TBX-Core module + TBX-Fiction module

NOTE This example demonstrates how the module name and dialect name can be different.

The data categories (and their accompanying constraints) included in the TBX-Fiction module are expressed in Table 1: