INTERNATIONAL STANDARD

ISO 30042

First edition
2008-12-15

Systems to manage terminology, knowledge and content — TermBase eXchange (TBX)

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

Reference number
ISO 30042:2008(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

Foreword

The International Organization for Standardization (ISO) 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 ISO 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 30042 was prepared by LISA OSCAR and was adopted, under a special "fast-track procedure", by Technical Committee ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 3, *Systems to manage terminology, knowledge and content*, in parallel with its approval by the ISO member bodies.

The Localization Industry Standards Association (LISA - www.lisa.org) is the standards organization for the globalization industry. Within LISA, the OSCAR (Open Standards for Container/content Allowing Reuse) Special Interest Group develops XML-based standards for automated language-processing in the areas of globalization, internationalization, localization, and translation, including standards for translation memory, terminology, text memory, word/character counts, and other related areas. The main task of the OSCAR Special Interest Group is to develop standards to facilitate and automate the globalization of products and services in a way that supports local language and culture conventions. Publication as an OSCAR standard requires approval by the OSCAR steering committee. An earlier version of TBX was developed and published by LISA in 2002.

TBX and the TBX logo are registered trademarks of LISA, and the TBX logo is subject to terms of use as defined by LISA. LISA maintains copyright on the TBX specification that is available on the LISA Web site, and ISO maintains copyright on the TBX specification that it distributes as ISO 30042. The technical content of these two documents is identical, and is subject to joint maintenance by a team of ISO TC 37 and LISA OSCAR members.

Introduction

This International Standard defines an XML-based framework for representing structured terminological data referred to as TermBase eXchange (TBX). Within this framework, a variety of terminological markup languages (TMLs) can be defined. A TML defined by TBX can facilitate the interchange of terminological data between users, which include people such as translators and writers, and 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. In this manner, it can enable the flow of terminological information throughout the information production cycle, both inside an organization and with outside service providers.

The intended audience for this document consists of two groups: (1) programmers and analysts who wish to develop software applications that process TBX-compliant data files; (2) terminologists and other language specialists who wish to analyse a terminological data collection for representation in TBX or to understand a TBX file.

This version of TBX is an update of a version that was published by the Localization Industry Standards Association (LISA) in 2002. Among other enhancements, the current version provides reference to an integrated schema that includes the core-structure module and the data-category constraints in combined declarations using the Relax NG and Schematron languages. It also provides reference to a TBX-compliant TML called TBX-Basic.

Users of this International Standard should first study the body (clauses 1-12). The suggested use of annexes A-I is described below.

(1) The core-structure module of TBX

All TMLs within the TBX framework have the same core structure. The core-structure module is described in Clause 8. A DTD for the core-structure module is found in Annex A. The elements, attributes, and data types are described in Annex D, and listed alphabetically in Annex I.

(2) The XCS module

TMLs may differ with respect to which data-categories are allowed, and at what levels of a terminological entry these data-categories can occur. These constraints on the core structure, which define a particular TML, are formally represented in an XCS file. A DTD for the XCS module is found in Annex B. The elements and attributes are described in Annex E, and listed alphabetically in Annex I.

(3) The default XCS of TBX

The TBX-default TML is constrained by the default XCS file. The TBX default XCS is described in Clause 9. The default XCS file is provided in Annex C. The data-categories are described in Annex D, and listed alphabetically in Annex I.

(4) Compliance checking of TBX document instances

Once a TBX TML has been defined by an XCS, a TBX document instance can be checked for compliance with that TML. The requirements for compliance are found in Clause 7. One can use a variety of methods and schema definition languages to check compliance. In particular, the Relax NG schema referred to in Annex F can be used to check whether a TBX document instance is compliant with the TBX-default TML. Annex F also indicates where a TBX user can find additional resources for compliance checking. Another TBX TML, called TBX-Basic, is referred to in Annex G.

(5) Changes that have been made to TBX since its submission to ISO in February 2007 are summarized in Annex H.

Summary of annexes:

A: DTD for core-structure module
B: DTD for XCS module
C: Default XCS that defines the TBX-default TML
D: Descriptions of core structure elements and attributes
D.5: Descriptions of default data-categories
E: Descriptions of XCS elements and attributes
F: Relax NG schema and other resources for compliance checking
G: Reference to TBX-Basic
H: Summary of changes to TBX
I: Indexes (alphabetical lists of elements and data-categories)

Systems to manage terminology, knowledge, and content - TermBase eXchange (TBX)

1 Scope

The TBX framework defined by this International Standard is designed to support various types of processes involving terminological data, including analysis, descriptive representation, dissemination, and interchange (exchange), in various computer environments. The primary purpose of TBX is for interchange of terminological data. It is limited in its ability to represent presentational markup. Intended application areas include translation and authoring.

TBX is modular in order to support the varying types of terminological data, or *data-categories*, that are included in different terminological databases (termbases). TBX includes two modules: a core structure, and a formalism for identifying a set of data-categories and their constraints, both expressed in XML. The term *TBX*, when used alone, refers to the framework consisting of these two interacting modules.

To maximize interoperability of the actual terminological data, TBX also provides a default set of data-categories that are commonly used in terminological databases. However, subsets or supersets of the default set of data-categories can be used within the TBX framework to support specific user requirements.

TBX, when used with its default set of data-categories, qualifies as a *terminological markup language (TML)* as defined in ISO 16642, which will be referred to as the *TBX-default TML* in this International Standard. Likewise, other markup languages that comply with TBX and use a subset of the default set of data-categories are also TMLs, but may go by other names, such as the one referred to in Annex G (Informative) TBX-Basic.

2 Normative references

The following referenced documents are indispensable for the application 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 639-1:2002, *Codes for the representation of names of languages – Part 1: Alpha-2 code*

ISO 639-2:1998, *Codes for the representation of names of languages – Part 2: Alpha-3 code*

ISO 639-3:2007, *Codes for the representation of names of languages – Part 3: Alpha-3 code for comprehensive coverage of languages*

ISO/IEC 646:1991, *Information technology – ISO 7-bit coded character set for information interchange*

ISO 3166-1:2006, *Codes for the representation of names of countries and their subdivisions – Part 1: Country codes*

ISO 8601:2004, *Data elements and interchange formats – Information interchange – Representation of dates and times*

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

ISO 12200:1999, *Computer applications in terminology – Machine-readable terminology interchange format (MARTIF) – Negotiated interchange*

ISO 12620, *Computer applications in terminology – Data categories*

ISO 16642:2003, *Computer applications in terminology – Terminological markup framework*

3 Terms and definitions

For the purposes of this International Standard, the following terms and definitions apply:

3.1
analysis
identification of the elements and structure of a terminological data collection in order to make explicit the data fields, their types, and their relationships

3.2
blindness
property of a data format indicating the degree to which the data are sufficiently defined that it is unnecessary for the importer to establish contact with the originator of the data in order to interpret them

NOTE The term blindness has its origin in the engineering phrase "blind transmission," which refers to a transmission of data where it is not necessary to “see” who is the sender of the data in order to interpret it. In terminology, the concept of blindness is often used in the context of **blind interchange** (3.3).

3.3
blind interchange
ability to receive a terminology file and integrate it into a target system, such as a Computer-Assisted Translation (CAT) tool, without having to contact the originator of the file in order to understand its contents

NOTE Interchange that is perfectly blind is interchange that is lossless while requiring no communication between the sender and the receiver of the data . Due to differences between terminological data collections and markup formats, perfectly blind interchange is rare. Typically, some of the data in a data collection is blind (can be exchanged without loss and without communication between parties) and some of the data requires communication between the parties in order to be exchanged.

3.4
complementary information
CI
information supplementary to that described in terminological entries and shared across the terminological data collection

[ISO 16642:2003]

NOTE In a TBX document instance, complementary information is contained in the back matter.

3.5
core-structure module
XML specification of the elements and attributes that are permitted in a TBX file

NOTE The core-structure module is defined in a DTD which is used in tandem with an XCS file that applies additional data-category constraints. It can also be used to generate an integrated schema, such as a Relax NG schema [ISO 19757-2], that defines both the core-structure module and the data-category constraints in one file. See also **data-category constraint** (3.7).

3.6
data-category
result of the specification of a given data field

[ISO 1087-2:2000]

EXAMPLE: /part of speech/, /grammatical number/

NOTE 1 The default set of data-categories for TBX were primarily selected from ISO 12620:1999.

NOTE 2 In running text, such as in this International Standard, data-category names are set off using forward slashes and italics. In a TBX document instance, camel case (e.g. <termNote type="partOfSpeech">noun</termNote>) should be used instead of using white space between words.

3.7
data-category constraint
specification of the value of an attribute, the content of an element, or one or more structural levels, that constrains the application of a **meta data-category** (3.16)

NOTE The data-category constraints are defined in an XCS file which is used in tandem with a DTD that defines the core-structure module. They can also be included in an integrated schema, such as a Relax NG schema, that incorporates both the core-structure module and the data-category constraints into one file. See also **core-structure module** (3.5).

3.8
data file
sequence of bytes that is either stored on a disc in a traditional file system or transmitted as a stream of data over a network

3.9
eXtensible Constraint Specification
XCS
XML file that identifies data-categories and their constraints for a specific **TBX TML** (3.21)

3.10
extension
totality of objects to which a concept corresponds

[ISO 1087-1:2000]

3.11
global information
GI
technical and administrative information applying to the entire data collection

[ISO 16642:2003]

NOTE In a TBX document instance, global information is contained in the front matter.

3.12
intension
set of characteristics which makes up the concept

[ISO 1087-1:2000]

3.13
interchange
exchange
transaction involving exporting data from one terminological data collection and importing it into another terminological data collection

3.14
lemma
lexical unit chosen according to lexicographical conventions to represent the different forms of an inflectional paradigm

[ISO 1951:2007]

3.15
lemmatize
to transform an inflected form of a word to its **lemma** (3.13)

3.16
lossless roundtrip
series of data manipulation procedures whereby data are output from a database into an interchange format and then re-imported into the same database without loss of information

3.17
meta data-category
name used to group similar **data-categories** (3.6) together; thus, a category of **data-categories** (3.6)

NOTE A meta data-category is equivalent to a typed element in ISO 16642. A meta data-category is instantiated into a terminological data-category through the value of its *type* attribute.

EXAMPLE: In the tag <descrip type="definition">, the meta data-category is *descrip* and the terminological data-category is /definition/.

3.18
metadata registry
information system for registering metadata

NOTE The associated information store or database is known as a metadata register.

3.19 object language
language being described in a <langSet>

EXAMPLE In a <langSet xml:lang='fr-FR'> element, the object language is French.

NOTE See also **working language** (3.28).

3.20
structural level
level of the metamodel to which one or more information units can be attached

[ISO 16642:2003]

3.21
TBX TML
TML (3.27) that adheres to **TBX** (3.26)

NOTE Implementers of a TBX TML may or may not use an XCS file (in conjunction with a DTD representing the core structure) for validation purposes. Some may choose to use an integrated schema instead.

3.22
TBX-default TML
TBX (3.26) and its default selection of **data-categories** (3.6) and their constraints expressed in the default **XCS** (3.9) file

3.23
TBX document instance
file containing terminological entries in a **TBX TML** (3.21) format

3.24
terminological database
database comprising information about special language concepts and terms designated to represent these concepts, along with associated conceptual, term-related, and administrative information

3.25
term component
one of the words comprising a multi-word term, or a component, such as a morpheme, of a single-word term

3.26
TermBase eXchange
TBX
framework consisting of a core structure, and a formalism (**eXtensible Constraint Specification** (3.9)) for identifying a set of **data-categories** (3.6) and their constraints, both expressed in XML

3.27
terminological markup language
TML
XML application for describing a terminological data collection conforming to the constraints expressed in ISO 16642 (Terminological markup framework)

NOTE 1 Adapted from ISO 16642:2003.

NOTE 2 TBX coupled with the default XCS file comprise a TML called **TBX-default TML** (3.22). TBX-Basic is also a TML.

3.28
working language
default language used in terminological entries

EXAMPLE If definitions, notes, picklist values, and so forth, are normally recorded in English, then English is the working language of the terminological data collection.

NOTE See also **object language** (3.19).

4 Relationship to other standards

The foundations for TBX were established by the following three international standards.

- ISO 16642:2003 (TMF) defines the structural metamodel for TBX and other TMLs
- ISO 12620 provides an inventory of data-categories for terminological data
- ISO 12200:1999 (MARTIF) provides the basis for the core structure of TBX and the XML styles of its elements and attributes.

A particular TML requires the choice of an XML style and a selection of data-categories. Most of the data-categories of the TBX-default TML were chosen from ISO 12620:1999, and the XML style of TBX was adopted from ISO 12200. Thus, TBX is a standards-based framework, being based on ISO 16642, ISO 12620, and ISO 12200.

5 Applications of TBX

TBX is designed to facilitate the following use cases:

Interchange, such as that required to support

- the flow of terminological data between technologies and systems
- integration of terminological data from multiple sources
- data conversion necessitated by a change in applications or technologies

Dissemination, including

- querying multiple terminological databases through a single user interface by passing data through a common intermediate format on a batch or dynamic basis
- placing data on an online site for download by interested parties
- making entries which require some work available for public feedback
- making terminology available dynamically in networked applications through a Web service

Analysis and representation, including

- comparing the contents of various terminological databases
- studying how lossless a conversion between two terminology databases can be
- designing a new terminological database intended to minimize loss during conversion.

6 Fundamental principles

6.1 General

The TBX framework is based on the assumption that, because of the variety of terminological data collections and use scenarios, no one terminological markup language would satisfy all user requirements.

To maximize interoperability, it is recommended that implementers of TBX adhere to ISO standards governing the principles and methodologies of terminology management, and the content and quality of terminological resources, such as those described in 2 Normative references and Bibliography. It is recommended that terminological databases select and use data-categories and their constraints that are specified in this International Standard. Extensions beyond those data-categories and constraints should be taken from ISO 12620 where possible. Fundamental principles of terminological data modelling such as data granularity, data elementarity, data repeatability, and term autonomy, are described in other ISO TC 37 standards.

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

TBX allows the representation of various kinds of information about individual terms that distinguish them from other terms in the same concept entry. It also allows for the documentation of directionality in situations where a term in one language may be translated by a given term in another language but the converse is not true due to partial equivalence. It should be noted that some terminological databases document nearly-identical concepts in separate linked entries, while others document nearly-identical concepts in the same entry. TBX can reflect both approaches.

6.2 Principles relating to grouping and representing data-categories

In TBX, there are four general types of data-categories. Understanding what these general types mean and how they are represented will facilitate the understanding of the rest of this International Standard.

NOTE In this specification, attribute names are identified by the *at* sign (@) in Appendix D, and are italicized in running text.

core-structure module data-category

A core-structure module data-category is any data-category that is defined in the core-structure module DTD. For example, <date>, <term>, and <descrip>.

meta data-category

A meta data-category is a core-structure module data-category that takes a *type* attribute, such as <descrip>, <admin>, and <termNote>. It is a *general* data-category that is used for grouping purposes and to reflect the metamodel in ISO 16642. Each *type* attribute value instantiates a meta data-category into a specific terminological data-category that is defined according to ISO 12620. The *type* attribute values are defined in an XCS file. For example, the tag <descrip type="definition"> comprises the meta data-category <descrip> instantiated into a terminological data-category that is called */definition/* according to ISO 12620.

data-category implemented using an attribute

A data-category implemented using an attribute is a terminological data-category that is defined according to ISO 12620, such as */definition/*, and one that is specified as a value of the *name* attribute in the default XCS file. In a TBX document instance, these data-categories appear as the value of a *type* attribute on a meta data-category element. The value of these data-categories is the content of their corresponding element. For instance, the */definition/* data-category, represented via the tag <descrip type="definition"> takes free text as its content, and the */gender/* data-category, represented in the tag <termNote type="grammaticalGender"> takes one of a closed set of values (picklist values) as its content (masculine, feminine, neuter, otherGender).

data-category implemented as the content of an element

A data-category implemented as the content of an element is a simple data-category, that is, one value of a closed set of values (picklist). These terminological data-categories are also documented according to ISO 12620. They are enumerated in the default XCS file as the permissible content of a meta data-category having a specific *type* attribute value. For instance, the meta data-category <termNote> that has the *type* attribute value 'termType' can have as its content a limited set of values that includes abbreviation, acronym, and so forth. In the integrated RNG schema that is referred to in Annex F, an element's content is constrained to a picklist through embedded Schematron rules. (For a reference to Schematron, see Bibliography.)

The use of meta data-categories in the TBX framework facilitates modularity. The core-structure (which remains

constant) is one module, and a particular XCS file (which expresses constraints on the core structure) is another module. The combination of these two modules defines a particular TML. This approach mirrors TMF (ISO 16642) in that the core-structure module corresponds to the abstract data model of TMF. In addition, it facilitates an explicit description of what two TMLs within the TBX framework have in common (the core structure) and how they differ (expressed as differences between their XCS files). This modular approach is consistent with generally accepted principles of modularity in software engineering, allowing a programmer/analyst to study the core structure and the XCS structure separately without being required to digest multiple large, monolithic schemas.

7 Requirements for TBX files

7.1 Compliance requirements

For a TML to be compliant with TBX, it shall meet the following three criteria:

1. The TML shall define XML document instances that are valid according to the TBX core-structure module. The core-structure module is described in 8 The core-structure module and is defined formally by the TBX DTD (Annex A).
2. The TML shall express its data-categories and their constraints in an XCS file that validates against the XCS DTD that is defined in Annex B (Normative) DTD for the data-category constraints (XCS file), and it shall adhere to the constraints in that XCS file. A TML that includes a data-category that has the same name as a data-category found in the default XCS shall use this data-category according to its description found in Annex C.
3. The TML can include fewer or more data-categories than those found in the default XCS (Annex C), and still be compliant with TBX, provided that those data-categories are expressed in an XCS file. If the TML includes data-categories that are not in the default XCS, it shall, in addition, describe those additional data-categories in the header of the XCS file.

NOTE Several general constraints, such as date formats, are not formally defined in either the DTD or the XCS, but are described in the relevant sections of this document, such as Annex D. These constraints shall also be adhered to for TBX compliance.

The extension for a TBX document instance is *.tbx* and the extension of an XCS document instance is *.xcs*.

Although an XCS document instance must exist to formally define a TBX TML, it need not be used for compliance checking. Indeed, general-purpose XML validation tools do not recognize the constraints in the XCS file unless those constraints have been incorporated into an integrated schema such as the one referred to in Annex F.

TBX compliance checking is schema-definition-language neutral. Three types of compliance checking are described in this document:

DTD and XCS

A DTD representing the core structure of TBX is provided at Annex A (Normative) DTD for the core structure module. An XCS file representing the default set of data-categories and their constraints is provided in Annex C.2 XCS file for the default data-categories and constraints. With a DTD and an XCS file, a TBX document instance can be validated by using a compliance checker that is specifically designed for TBX files.

Relax NG

A Relax NG schema file representing the core structure and the default set of data-categories and their constraints is referred to in Annex F. This file includes some embedded Schematron for some of the data-category constraints. By using this file, one can validate a TBX document instance for compliance with the default TBX TML by using any XML validator that supports Relax NG and Schematron. With appropriate software, an integrated Relax NG schema could be generated for another TBX TML, based on its XCS.

Other methods

Compliance checking is also allowed using other methods that incorporate information from the core-structure module and data-category constraints. Additional methods may be documented on the LISA Web.

For more information about user-specific TMLs based on TBX, see 12 Creating customized TBX TMLs.

7.2 Examples of non-compliance

Compliance to TBX includes the following aspects:

1. XML well-formedness
2. validity relative to the core-structure module
3. adherence to the data-category constraints in an XCS file

The following example is not well-formed, since the first <descrip> element has a spelling error in the end tag and the <term> element has no closing tag.

```
<term>kitten <descrip type='definition'>Small feline</deescrip>
```

The following example is well-formed but not core-structure valid, since the core-structure module of TBX does not allow a <descrip> tag to follow a <tig>.

```
<langSet xml:lang="fr-ca">
<tig><term>zone de soufflage</term>
</tig>
<descrip type="definition">Area where snow is thrown by a snowplow.</descrip>
</langSet>
```

The following example is valid according to the TBX core-structure DTD but does not adhere to the default XCS, since there is no TBX data-category called "conflagration" in the XCS file.

```
<term>kitten</term> <descrip type='conflagration'>Small feline</descrip>
```


7.3 Implementation levels

There are three levels of implementation of TBX for a given software application relative to a particular terminological database:

Level 1
The software application shall export and import TBX document instance files that are well-formed and core-structure valid and that adhere to at least one XCS file, and the software application shall detect when document instances are not well-formed or not core-structure valid or not XCS-adherent. The XCS file is not required to be the default XCS, for example, it could be a superset or a subset. Level 1 supports interchange between systems that use the same XCS.

Level 2
The software application shall achieve level one implementation but shall also be able to import every data-category that is in the default XCS. Thus, level two implementation supports a degree of blindness in that it can import TBX files from any outside source whose export can be limited to the data-categories that conform to the default XCS file.

Level 3
The software application shall achieve level two implementation and in addition be able to check adherence to a comprehensive XCS that supports a lossless roundtrip from the terminological database in the application to a TBX TML and back to the terminological database in the application. Thus, once the information in the terminological database has been exported to TBX, the terminological database can be emptied and subsequently repopulated from the information in the TBX file.

8 The core-structure module

8.1 Introduction

This section describes the core-structure module for TBX. The elements of the core-structure module are formally declared in Annex A and described in Annex D. For quick access to all these elements, refer to the Index (Annex I).

There is a correspondence between the high-level elements of the core-structure module and the TMF (ISO 16642) metamodel, shown in Figure 1, High-level structure of the TMF (ISO 16642) metamodel. The Terminological Data Collection (TDC) corresponds to a TBX document instance.

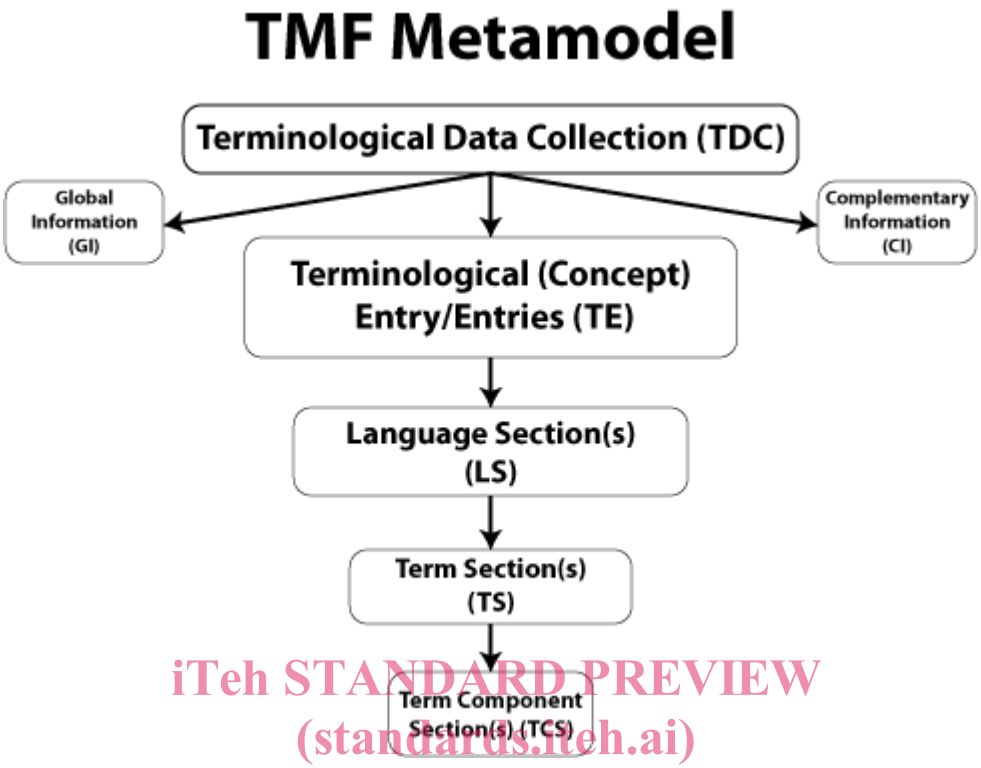

Figure 1. High-level structure of the TMF (ISO 16642) metamodel

In the figures in the following sections, a question mark after an element in the box-and-line diagrams indicates that a single occurrence of the element is optional. A plus sign after an element indicates that one or more occurrences of the element are required. A plus sign in a box by itself indicates that the structure connected to the right of the box occurs one or more times. An asterisk indicates that an element is optional and that multiple instances are allowed. A dotted line indicates a logical exclusive OR relationship, i.e., either of two elements must occur.

8.2 Hierarchy

The highest-level XML element in a TBX document instance is the <martif> element, which consists of a <martifHeader> element and a <text> element. (See Figure 2, The MARTIF structure.) These element names are taken from ISO 12200 and have roots in the Text Encoding Initiative.

The <text> element in Figure 2 consists of terminological entries, which are enclosed within one <body> element, and complementary information (a metamodel object class). In TBX, complementary information is found in the <back> element.

The <martifHeader> element corresponds to *global information* in the TMF metamodel and consists of a description of the whole terminological data collection (in the <fileDesc> element), information about the applicable XCS file (in the <encodingDesc> element), and a history of major revisions to the collection (in the <revisionDesc> element). Character encoding information shall be included in the header only when the encoding attribute of the XML declaration in the TBX document instance is not a Unicode encoding value.