INTERNATIONAL STANDARD

ISO 16642

Second edition
2017-11

Computer applications in terminology — Terminological markup framework

Applications informatiques en terminologie — Plate-forme pour le balisage de terminologies informatisées

Reference number
ISO 16642:2017(E)

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

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, *Terminology and other language and content resources*, Subcommittee SC 3, *Computer applications for terminology*.

This second edition cancels and replaces the first edition (ISO 16642:2003), which has been technically revised.

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

— The following formats are no longer actively used. Consequently, references to these formats have been removed (including Annex A, Annex B, and Annex C):

  — Martif with specified constraints (MSC);

  — Geneter;

  — Data category interchange format (DCIF);

  — Generic mapping tool (GMT).

— With the removal of Annex B and Annex C, this document no longer includes any comprehensive code examples of a TML. Examples of TMLs are now available in ISO 30042, TermBase eXchange, and also at the following Web site: www.tbxinfo.net.

— References to the former ISO/TC 37 Data Category Registry or ISOcat have been changed from normative to informative. In addition, the name has changed to DatCatInfo, now as an example of data category repositories.

— References to ISO 12620:1999 and ISO 12620:2009 have been removed. These previous standards have been withdrawn.

— The TypedValuedElement style has been added.

— Examples have been updated to reflect ISO 30042:2008 (TBX). TBX-Basic is mentioned as a TML.

— Some of the examples and tables have been moved to appropriate sections.

— As a consequence of the aforementioned changes, some historical, didactic, or duplicate information has been removed to adhere more closely to ISO editorial standards.

Introduction

Terminological data are collected, managed and stored in a wide variety of systems, typically various kinds of database management systems, ranging from personal computer applications for individual users to large terminological database systems operated by major companies and governmental agencies. Terminology databases are comprised of various types of information, called data categories, and can adopt different structural models. However, terminological data often need to be shared and reused in a number of applications, and this sharing is facilitated when the data adheres to a common model. To facilitate co-operation and to prevent duplicate work, it is important to develop standards and guidelines for creating and using terminological data collections (TDCs) as well as for sharing and exchanging data.

This document presents a modular approach for analysing existing TDCs and designing new ones. It also provides a framework for defining terminological markup languages (TMLs) that are interoperable.

This document makes reference to DatCatInfo, an example of an available data category repository. DatCatInfo is an online database of information about the types of data that can be included in terminological data collections and other language resources. It is available at www.datcatinfo.net.

Computer applications in terminology — Terminological markup framework

1 Scope

This document specifies a framework for representing data recorded in terminological data collections (TDCs). This framework includes a metamodel and methods for describing specific terminological markup languages (TMLs) expressed in XML. The mechanisms for implementing constraints in a TML are defined, but not the specific constraints for individual TMLs.

This document is designed to support the development and use of computer applications for terminological data and the exchange of such data between different applications. This document also defines the conditions that allow the data expressed in one TML to be mapped onto another TML.

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 704, *Terminology work — Principles and methods*

ISO 1087-1, *Terminology work — Vocabulary — Part 1: Theory and application*

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

ISO 26162, *Systems to manage terminology, knowledge and content — Design, implementation and maintenance of terminology management systems*

ISO 30042:2008, *Systems to manage terminology, knowledge and content — TermBase eXchange (TBX)*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 1087-1 and the following apply.

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

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

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

3.1
basic information unit
information unit (3.12) attached to a *component* (3.3) of the metamodel and that can be expressed by means of a single *data category* (3.6)

3.2
complementary information
CI
information supplementary to that described in *terminological entries* (3.22) and shared across the *terminological data collection* (3.21)

Note 1 to entry: Domain hierarchies, institution descriptions, bibliographic references and references to text corpora are typical examples of complementary information.

3.3
component
elementary description unit of a metamodel to which *data categories* (3.6) can be associated to form a data model

3.4
compound information unit
information unit (3.12) attached to a *component* (3.3) of the metamodel that is expressed by means of several grouped *data categories* (3.6), that, taken together, express a coherent unit of information

3.5
conceptual domain
set of valid value meanings associated with a *data category* (3.6)

Note 1 to entry: For example, the data category /part of speech/ could have the following conceptual domain: /noun/, /verb/, /adjective/, /adverb/, and so forth.

3.6
data category
elementary descriptor used in a linguistic description or annotation scheme

Note 1 to entry: In this document, data categories are indicated in between forward slashes (/), e.g. /definition/.

3.7
data category repository
DCR
electronic repository of *data category specifications* (3.9) to be used as a reference for the definition of linguistic annotation schemes or any other representation model for language resources

Note 1 to entry: A DCR for language resources is available at http://www.datcatinfo.net.

3.8
data category selection
DCS
set of *data categories* (3.6) selected from a *DCR* (3.7)

3.9
data category specification
set of attributes used to fully describe a given *data category* (3.6)

Note 1 to entry: The abbreviation "DCS" is associated with data category selection and is not used for data category specification.

3.10
expansion tree
structured group of XML elements that implement a level of the metamodel in a given *TML* (3.23)

3.11
global information
GI
technical and administrative information applying to the entire *terminological data collection* (3.21)

Note 1 to entry: For example, the title of the terminological data collection, revision history, owner or copyright information.

3.12
information unit
IU
elementary piece of information attached to a structural level of the metamodel

3.13
language section
LS
part of a *terminological entry* (3.22) containing information related to one language

Note 1 to entry: One terminological entry may contain information on one or more languages.

3.14
object language
language being described

3.15
persistent identifier
PID
unique Uniform Resource Identifier (URI) that assures permanent access for a digital object by providing access to it independently of its physical location or current ownership

3.16
structural node
instance of *component* (3.3) within the representation of a *terminological data collection* (3.21)

3.17
structural skeleton
abstract description of an instance of a *terminological data collection* (3.21) in conformity with the metamodel

3.18
style
specification for the implementation of a *data category* (3.6) in XML

3.19
term component section
TCS
part of a *term section* (3.20) giving linguistic information about the components of a term

3.20
term section
TS
part of a *language section* (3.13) giving information about a term

3.21
terminological data collection
TDC
resource consisting of *terminological entries* (3.22) with associated meta data and documentary information

3.22
terminological entry
TE
part of a *terminological data collection* (3.21) which contains the terminological data related to one concept

Note 1 to entry: Every element in the TE can be linked to complementary information, to other terminological entries and to other elements in the same terminological entry.

3.23
terminological markup language
TML
XML format for representing a *terminological data collection* (3.21) conforming to the constraints expressed in this document

3.24
Unified Modeling Language
UML
language for specifying, visualizing, constructing and documenting the artifacts of software systems

3.25
vocabulary
<data modeling> set of strings used to implement a *data category* (3.6) according to a *style* (3.18)

3.26
working language
language used to describe objects

3.27
XML outline
part of a *terminological data collection* (3.21) corresponding to the XML implementation of the metamodel

4 Modular approach

Terminological Markup Framework (TMF) consists of two levels of abstraction. The first (and most abstract) level is the metamodel level. The metamodel level supports analysis, design and exchange at a very general level, i.e. it is independent of any specific implementation or software. The metamodel shall be shared by all TDCs that are compliant with TMF. The second level is the data model level, which adds the necessary data categories for representing specific TDCs.

The implementation of a data model in XML is called a terminological markup language (TML). TMLs can be described on the basis of a limited number of characteristics, namely

— how the TML expresses the structural organization of the metamodel (i.e. the expansion trees of the TML);

— the specific data categories used by the TML and how they relate to the metamodel;

— the way in which these data categories can be expressed in XML and anchored on the expansion trees of the TML, i.e. the XML style of any given data category;

— the vocabularies used by the TML to express those various informational objects as XML elements and attributes according to the corresponding XML styles.

Figure 1 represents the information required to fully specify a TML.

— The metamodel describes the basic hierarchy of components to which any TML shall conform.

— A set of data category specifications from a data category repository, which can form the basis for defining a data category selection (DCS) for the TML

— The dialectal specification (dialect) includes the various elements needed to represent a given TML in an XML format. These elements comprise expansion trees and data category instantiation styles, together with their corresponding vocabularies.

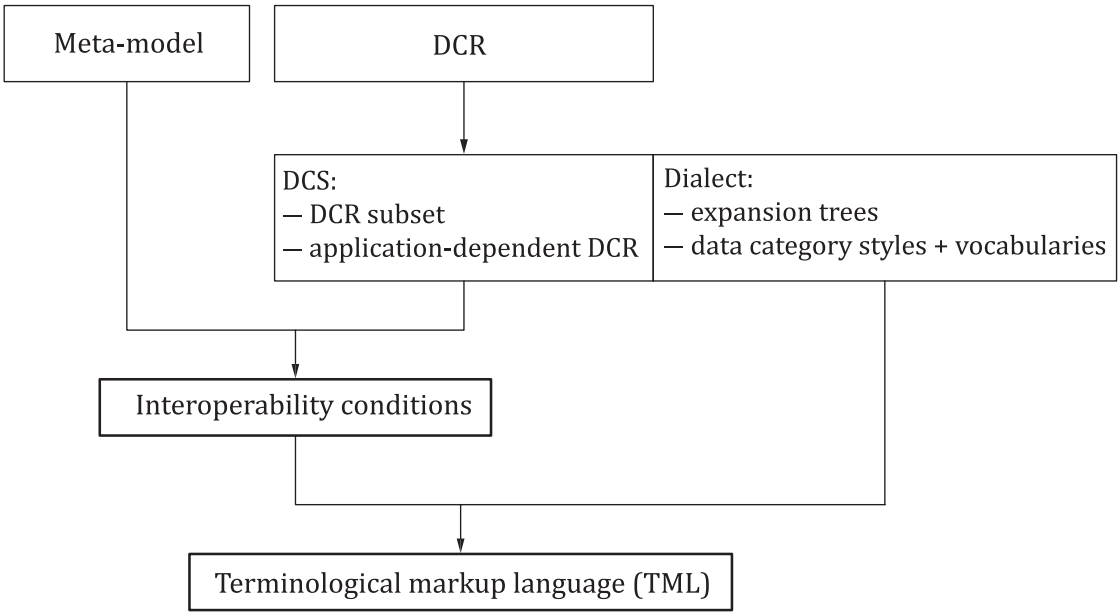

Figure 1 — Various knowledge sources involved in the description of a TML

A DCR providing sample data category specifications for language resources is available at www.datcatinfo.net. Where possible, data categories documented in this DCR should be used for a TML. If no suitable data category is available in this DCR, the implementers of the TML should propose the creation of the required data category specification within this DCR.

5 Generic model for describing terminological data

5.1 Principles

This clause describes a class of XML document structures which can be used to represent a wide range of terminological data formats, and provides a framework for representing these document structures in XML.

Each type of document structure is described by means of a three-tiered information structure that describes:

— a *metamodel*, which comprises a hierarchy of components;

— *information units*, which can be associated with each component of the metamodel;

— *annotations*, which can be used to qualify properties associated with a given information unit.

Information units can be basic or compound. A basic information unit encapsulates information that can be expressed by means of a single data category. A compound information unit encapsulates information that is expressed by means of several grouped data categories that, taken together, express a coherent unit of information. For instance, a compound information unit can be used to represent the fact that a transaction can be a combination of a transaction type (such as modification), the person who performed it, and the date when it was performed.

Basic information units, whether they are directly attached to a component or are placed within a compound information unit, can take two non-exclusive types of value:

— an atomic value corresponding either to a simple type (in the sense of XML schemas) such as a number, string, element of a picklist, etc., or to a mixed content type in the case of annotated text;

— a reference to a component in order to express a relation between it and the current component.