

Computer applications in terminology — Terminological markup framework

ICS 01.020; 35.240.30

National foreword

This British Standard reproduces verbatim ISO 16642:2003 and implements it as the UK national standard.

The UK participation in its preparation was entrusted to Technical Committee TS/1, Terminology and other language resources, which has the responsibility to:

- aid enquirers to understand the text;
- present to the responsible international/European committee any enquiries on the interpretation, or proposals for change, and keep the UK interests informed;
- monitor related international and European developments and promulgate them in the UK.

A list of organizations represented on this committee can be obtained on request to its secretary.

Cross-references

The British Standards which implement international publications referred to in this document may be found in the *BSI Catalogue* under the section entitled “International Standards Correspondence Index”, or by using the “Search” facility of the *BSI Electronic Catalogue* or of British Standards Online.

This publication does not purport to include all the necessary provisions of a contract. Users are responsible for its correct application.

Compliance with a British Standard does not of itself confer immunity from legal obligations.

Summary of pages

This document comprises a front cover, an inside front cover, the ISO title page, pages ii to vi, pages 1 to 49 and a back cover.

The BSI copyright notice displayed in this document indicates when the document was last issued.

Amendments issued since publication

Amd. No.	Date	Comments

This British Standard was published under the authority of the Standards Policy and Strategy Committee on 15 September 2003

© BSI 15 September 2003

ISBN 0 580 42636 X

INTERNATIONAL STANDARD

BS ISO 16642:2003

ISO
16642

First edition
2003-08-15

Computer applications in terminology — Terminological markup framework

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



Reference number
ISO 16642:2003(E)
Documents
Document Sharing Hub

Contents

Page

1	Scope	1
2	Normative references	1
3	Terms and definitions	1
4	General principles and interoperability principle	4
5	Generic model for describing linguistic data and its application to terminology	6
5.1	Introduction	6
5.1.1	General principles	6
5.1.2	Example	7
5.2	Generic representation of structural levels and information units	8
5.3	The terminological meta-model	9
5.4	Designing representations of terminological data on the basis of the meta-model	12
5.5	Interchange, dissemination and interoperability	12
5.6	XML canonical representation of the generic model	13
5.6.1	Introduction	13
5.6.2	Example	13
5.6.3	Description of the GMT format	14
5.7	Representing languages in a terminological data collection	17
6	Defining a TML	18
6.1	General	18
6.2	Defining interoperability conditions	18
6.3	Implementing a TML	18
6.3.1	Introduction	18
6.3.2	Implementing the meta-model	18
6.3.3	Anchoring data categories on the TML XML outline	19
6.3.4	Implementing annotations	20
6.3.5	Implementing brackets	21
6.3.6	Namespaces	21
Annex A (normative)	XML schema of the GMT format	22
Annex B (normative)	The MSC TML	24
Annex C (normative)	The Geneter TML	29
Annex D (informative)	Conformance of terminological data to TMF	43
Bibliography		48

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 16642 was prepared by Technical Committee ISO/TC 37, *Terminology and other language resources*, Subcommittee SC 3, *Computer applications for terminology*.

Introduction

Terminological data are collected, managed and stored in a wide variety of systems, typically in applications, i.e. various kinds of database management system, ranging from personal computer applications for individual users to mainframe term-bank systems operated by major companies and governmental agencies. Termbases are comprised of various sets of data category and are based on various kinds of data model. Terminological data often need to be shared and reused in a number of applications, and this sharing is usually accomplished using intermediate formats. To facilitate co-operation and to prevent duplicate work, it is important to develop standards and guidelines for creating and using terminological data collections as well as for sharing and exchanging data.

The meta-model defined in this International Standard fits into an integrated approach to be used in analysing existing terminological data collections and in designing new ones, which are typically processed using relational or text-based data management systems. Terminological data collections can also be stored as structured documents with markup based on formats that are typically defined using Standard Generalized Markup Language (SGML), defined in ISO 8879 [12], or eXtensible Markup Language (XML), which is based on SGML but amended for use on the World Wide Web by the World Wide Web Consortium (W3C). An integrated approach eases the tasks of importing data from a flat file with markup into a database and of exporting data from a database to a structured document. Another motivation for an integrated approach, as opposed to entirely separate approaches for databases and structured documents, is that XML-based formats are now being processed in new ways, similar to traditional database management systems. For example, XML files are being queried and updated directly without importing data into traditional database environments.

This integrated approach to analysis and design consists of two levels of abstraction. The first (and most abstract) level of the integrated approach is the meta-model level. The meta-model level, which could also be called the abstract conceptual data model level, supports analysis and design at a very general level. The second level is the data model level.

At the data model level, the designer of the terminological data collection has the possibility to make various choices, based on real-life needs. First, designers must determine the form of representation most appropriate for their terminological data, addressing the following choices:

- whether to use a relational database or a flat file with markup;
- whether the data will be used primarily for queries and updates, and be represented in some database management system and, if this is the case, what system to use;
- whether the data will be used primarily for sharing and interchange, and be represented in a flat file with markup.

For the purposes of this International Standard it is assumed that all flat files will use XML markup.

Once the choice between a database management system and a flat file with XML markup has been made, a data model must be chosen. For a relational database, a typical method of describing a data model is an entity-relationship diagram. For an XML document, a typical method of describing a data model is a Document Type Definition (DTD). An alternative method, using what is called an “XML schema”, is provided by the W3C. In the future, it will be possible to use more abstract methods of describing an XML format.

A specific implementation of the meta-model for terminology markup expressed in XML is called a terminological markup language (or TML), which can be described on the basis of a limited number of characteristics, namely

- how the TML expresses the structural organization of the meta-model (i.e. the expansion trees of the TML),
- the specific data categories used by the TML and how they relate to the meta-model,

- the way in which these data categories can be expressed in XML and thus anchored on the expansion trees of the TML, i.e. the XML style of any given data category, and
- the vocabularies used by the TML to express those various informational objects as XML elements and attributes according to the corresponding XML styles.

Some of the examples in this International Standard are instances of the MSC (MARTIF with Specified Constraints) and Geneter formats as described in Annex B and Annex C respectively.

Computer applications in terminology — Terminological markup framework

1 Scope

This International Standard specifies a framework designed to provide guidance on the basic principles for representing data recorded in terminological data collections. This framework includes a meta-model and methods for describing specific terminological markup languages (TMLs) expressed in XML. The mechanisms for implementing constraints in a TML are defined in this International Standard, but not the specific constraints for individual TMLs, except for the three TMLs defined in Annexes B to D.

This International Standard is designed to support the development and use of computer applications for terminological data and the exchange of such data between different applications. It does not standardize data categories and methods for the specification of data structures which are specified in ISO 12620 and other related International Standards.

This International Standard also defines the conditions that allow the data expressed in one TML to be mapped onto another TML and specifies a generic mapping tool (GMT) for this purpose (see Annex A).

In addition, this International Standard describes a generic model for describing linguistic data.

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

ISO 1087-2, *Terminology work — Vocabulary — Part 2: Computer applications*

ISO 12620:1999, *Computer applications in terminology — Data categories*

Extensible Markup Language (XML) 1.0, Second edition, BRAY, T., PAOLI, J., SPERBERG-MCQUEEN, C. M., and MALER, E. (eds.), W3C Recommendation 6 October 2000, available at <<http://www.w3.org/TR/REC-xml>>

Dublin Core Qualifiers, 2000-07-11, available at <<http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/>>

XHTML™ 1.0 The Extensible HyperText Markup Language, 2nd edition, available at <<http://www.w3.org/TR/xhtml1/>>

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 1087-1 and ISO 1087-2, and the following apply.

3.1

CI

complementary information

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

NOTE Domain hierarchies, institution descriptions and bibliographical references are typical examples of complementary information.

3.2

data category

result of the specification of a given data field

[ISO 1087-2:2000, definition 6.14]

NOTE 1 A data category is a type of data field, such as Definition.

NOTE 2 ISO 12620 is an inventory of data categories, i.e. a “DCR” (see 3.3).

3.3

DCR

data category registry

data category specification used as a normative reference for the description of a TML

NOTE ISO 12620:1999 is a typical DCR in the context of this International Standard.

3.4

DCS

data category selection

component of a TML's specification that constrains its informational content

NOTE The informational content may be constrained, for example by specifying which data categories are allowed and how each data category can be used.

3.5

expansion tree

list of XML elements together with their organization that implement a level of the meta-model in a given TML

3.6

GMT

generic mapping tool

canonical representation of the terminological markup framework model in XML

3.7

GI

global information

technical and administrative information applying to the entire data collection

EXAMPLE Title of the data collection, revision history.

3.8

information unit

IU

elementary piece of information attached to a level of the meta-model

3.9

LS

language section

part of a terminological entry containing information related to one language

NOTE One terminological entry may contain information on one, two or more languages.

3.10**object language**

language being described

3.11**structural level**

level of the meta-model to which one or more information units can be attached

3.12**structural skeleton**

abstract description of an instance of a terminological database in conformity with the meta-model

3.13**style**

properties relating to a data category that determine how it may be expressed in XML

3.14**TCS****term component section**

part of a term section giving linguistic information about the components of a term

3.15**TS****term section**

part of a language section giving information about a term

EXAMPLE Usage of a term, term elements.

3.16**TDC****terminological data collection**

collection of data containing information on concepts of specific subject fields

[ISO 1087-2:2000, definition 2.21]

NOTE For the purposes of this International Standard, terminological data collections are assumed to contain GI and CI in addition to strictly terminological information.

3.17**TE****terminological entry**

entry containing information on terminological units

EXAMPLE Subject-specific concepts, terms, etc.

NOTE Every element in the TE can be linked to CI, to other entries and to other elements in the same entry.

3.18**TML****terminological markup language**

XML application for describing a TDC conforming to the constraints expressed in this International Standard

3.19**UML****unified modelling language**

language for specifying, visualizing, constructing and documenting the artefacts of software systems

3.20**vocabulary**

set of strings used to implement a data category according to a style

3.21

working language

language used to describe objects

3.22

XML outline

part of a terminological database corresponding to the XML implementation of the meta-model

4 General principles and interoperability principle

Describing a specific TML can be seen as a process involving several knowledge sources which interact with one another at various levels. This process leads to the required specification of two important aspects of a TML:

- the informational properties of the TML, i.e. its capacity to represent a given piece of information related to the terminological description;
- the way the TML can be expressed, for instance as an XML document.

Figure 1 represents the various knowledge sources that form the basis of this International Standard and that can lead to the full specification of a TML.

Two of those knowledge sources are shared by all TMLs and can be seen as reference material for this International Standard.

- The meta-model describes the basic hierarchy of structural levels to which any TML shall conform as defined in this International Standard.
- A DCR is a set of data category specifications on which any specific TML shall rely for creating its own data category set. For the application of this International Standard, ISO 12620 forms a reference DCR for any information unit to be used in the specification of a TML.

Two other knowledge sources are used to define the specific information units of a given TML from the point of view of both its informational properties and its representation in XML.

- The DCS describes the set of data categories that can be used within a given TML. The DCS can comprise both a subset of the DCR together with any idiosyncratic data categories needed for a specific application.
- The dialectal specification (Dialect) includes the various elements needed to describe a given TML as an XML document. These elements comprise expansion trees and data category instantiation styles, together with their corresponding vocabularies.

The combination of the meta-model and a given DCS is enough to define conditions of interoperability, encompassing the full informational properties of the TML from a terminological point of view. Any information structure that corresponds to such conditions has a canonical expression as an XML document using the GMT representation. The interoperability between two different TMLs depends solely on their compatibility at that level (see Figure 2).

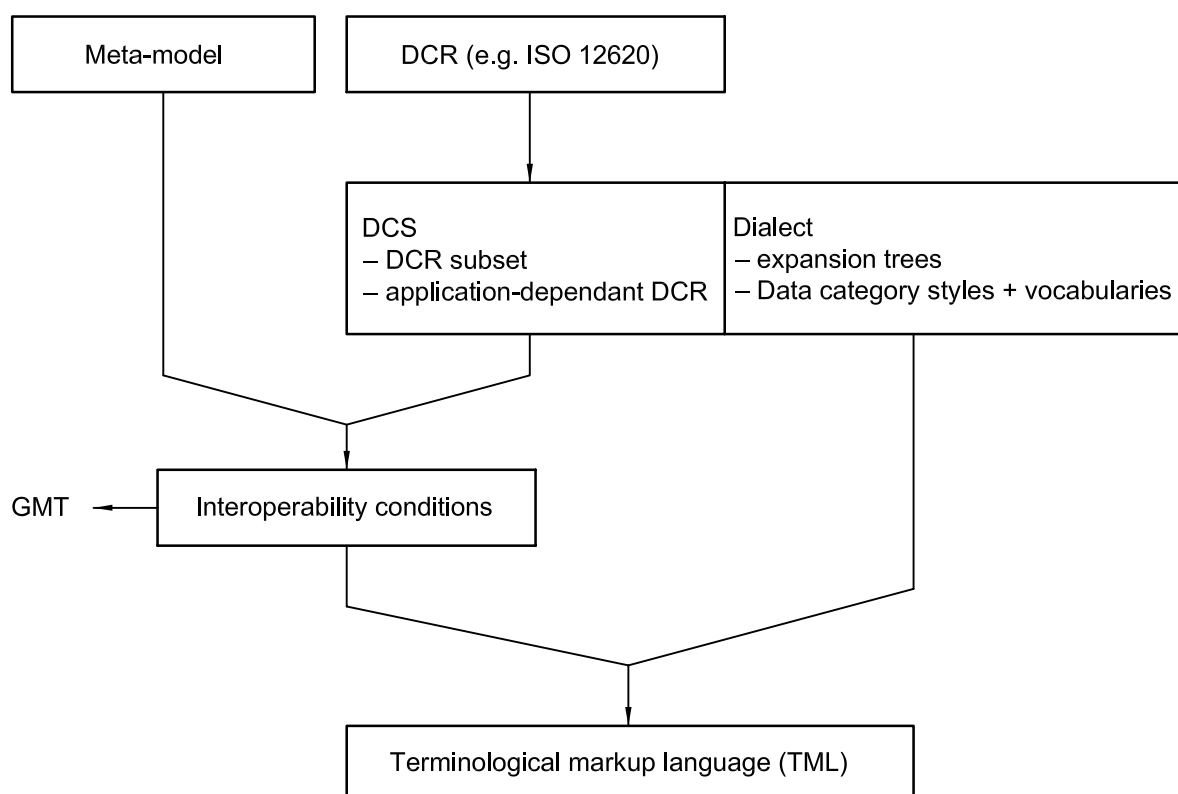


Figure 1 — The various knowledge sources involved in the description of a TML

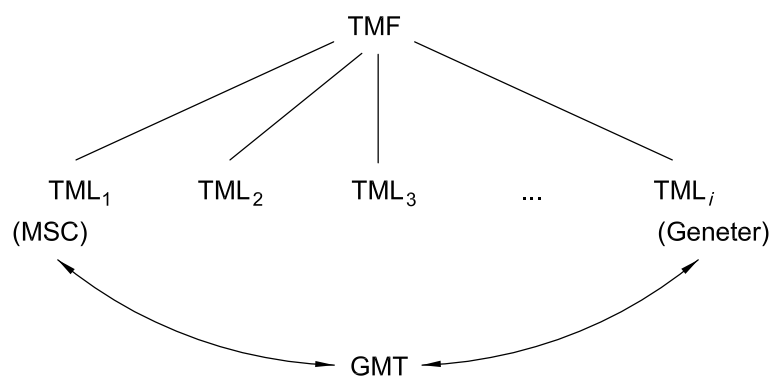


Figure 2 — Interoperability between two TMLs using the GMT

More precisely, the interoperability between two TMLs can be defined by comparing their specifications. Under the condition that two TMLs are based upon the same DCS, any terminological data collection expressed in either TML can be transformed, without any loss of information, into a collection represented using the other format. In particular, the additional specifications that are needed to express interoperability conditions in the

form of an XML implementation of a TML do not influence the level of interoperability between two markup languages. The preceding interoperability principle leads to two immediate consequences:

- it justifies the role of the GMT as a pivot markup language when transforming any data expressed in a TML into data expressed in another TML;
- when two TMLs are based upon two different DCSs, it provides a framework for identifying what information can be transformed between one format and another, and what will be lost during the transformation (weak interoperability condition).

The combination of interoperability conditions with dialectal information provides the information necessary to express a given TML in XML.

5 Generic model for describing linguistic data and its application to terminology

5.1 Introduction

5.1.1 General principles

Linguistic structures exist in a wide variety of formats ranging from highly organized data (e.g. terminological data bases) to loosely structured information (e.g. text lightly annotated for names, dates, etc.). The representation of annotated corpora is not hierarchical, but based on the expression of multiple views representing various levels of linguistic information, usually pointing to primary data (e.g. part of speech tagging) and sometimes to one another (e.g. reference annotation based on basic phrase structure annotation).

The following model identifies a class of document structures which could be used to cover a wide range of linguistic annotation formats, and provides a framework which can be applied using XML.

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

- a *meta-model*, which represents a hierarchy of structural nodes which are relevant for linguistic description;
- specific *information units*, which can be associated with each structural node of the meta-model;
- relevant *annotations*, which can be used to qualify some parts of the value associated with a given information unit.

Each structural node can be qualified by a group of basic or compound information units. A basic information unit describes a property that can be directly expressed by means of a data category. A compound information unit corresponds to the grouping at one level of several basic information units, which, 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, a responsibility and the transaction date.

Basic information units, whether they are directly attached to a structural node or 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 structural node in order to express a relation between it and the current structural node.

Basic and compound information units can be abstractly represented as feature-value structures associated with specific structural nodes in the structural skeleton. For instance, a Geneter sub-document identified as `<Owner>UHB</Owner>` can be modelled as a basic information unit in the following way:

```
[owner = UHB]
```

Similarly, the following MSC sub-document

```
<transacGrp>
  <transac>modification</transac>
  <transacNote type="responsiblePerson">YYY</transacNote>
  <date>1964-4-4</date>
</transacGrp>
```

can be modelled as

$$\left[\text{transacGrp} = \left[\begin{array}{l} \text{transac} = \text{modification} \\ \text{responsiblePerson} = \text{YYY} \\ \text{date} = 1964-4-4 \end{array} \right] \right]$$

The preceding model for information units is to be completed by a last level of information representation, which corresponds to the association of semantic information to subparts of information unit values. Such *annotations* typically occur when one wants to identify, within a terminological definition, specific references to information involving genus and/or differentia. See, for instance the following definition for lead pencil:

```
<definition>
  <broaderConcept>pencil</broaderConcept> whose
  <characteristic>casing</characteristic> is fixed around a central
  <characteristic>graphite</characteristic> medium which is
  <characteristic>used for writing or making marks</characteristic>
</definition>
```

Such information, also known as *mixed content* in XML, cannot be directly represented as a feature structure and will be directly expressed in the following GMT representation (<annot> element).

5.1.2 Example

To illustrate how a TDC can be analysed as an abstract structure, let us consider a simple terminological entry expressed as an XML document conforming to MSC specifications:

```
<?xml version="1.0"?>
<martif type="MSC" lang="en">
  <text>
    <body>
      <termEntry id="ID67">
        <descrip type="subjectField">manufacturing</descrip>
        <descrip type="definition">A value between 0 and 1 used in ...
        </descrip>
        <langSet lang="en">
          <tig>
            <term>alpha smoothing factor</term>
            <termNote type="termType">fullForm
            </termNote>
          </tig>
        </langSet>
        <langSet lang="hu">
          <tig><term>Alfa ...</term></tig>
        </langSet>
      </termEntry>
    </body>
  </text>
</martif>
```

The XML document represented above can be mapped to the abstract model described in this clause by identifying a structural skeleton corresponding to the meta-model and by associating the corresponding information units with each structural node in the structural skeleton, as shown in Figure 3.

Here, data categories can be mapped onto the corresponding data categories specified in ISO 12620:

Data category	ISO 12620:1999 number	ISO 12620:1999 name
id	A.10.15	entry identifier
subjectField	A.4	subject field
definition	A.5.1	definition
lang	A.10.7.1	language identifier
term	A.1	term
termType	A.2.1	term type
fullForm	A.2.1.7	full form

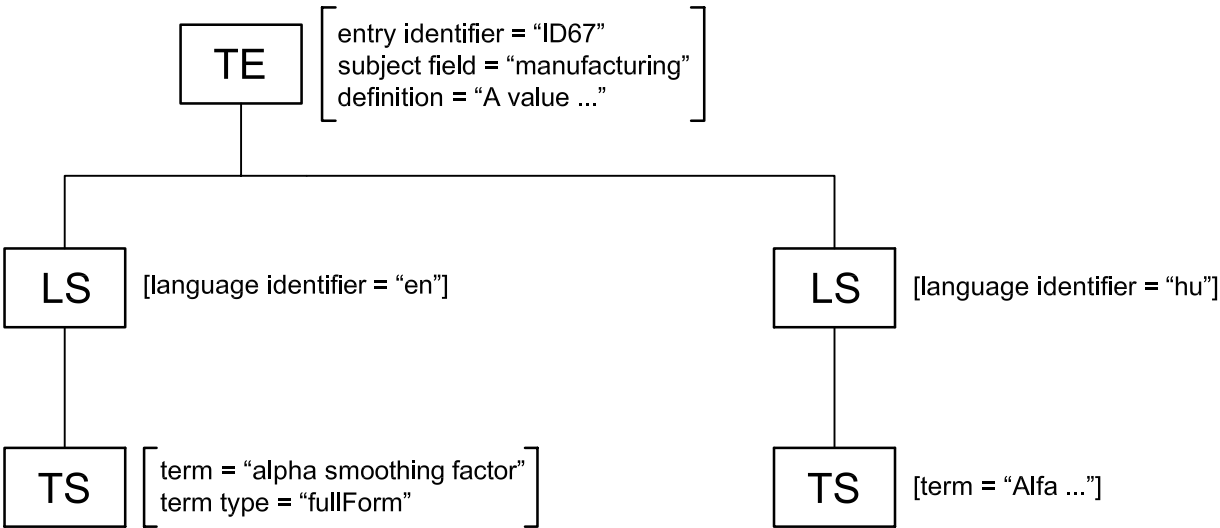


Figure 3 — Mapping an XML document to the abstract model

5.2 Generic representation of structural levels and information units

5.2.1 Linguistic data can be represented using a generic architecture that consists of a graph of elementary structural nodes to which one or more information units are attached. This architecture is shown in a UML diagram in Figure 4. The diagram expresses the relationship between the following defined classes:

- structural node: a class containing one attribute (LevelName) that identifies objects of this type in the context of a given Linguistic Resource (LR) format (e.g. TE/Terminological Entry for the representation of TermBanks);
- information unit: a class containing three attributes to identify objects of this type in relation to a given data category (IUName, e.g. Definition, PartOfSpeech, etc.), to describe a type for its content (C_type) and to provide the actual content value (C_value).

NOTE The value of C_type can either belong to the set of simple types as defined in [XML Schema Part 2: Datatypes](#) or be MIXED as described in 5.2.2.

Objects of these two classes can be related in the following ways.

- association: Indicates that a structural node is related to another structural node by a hierarchical link. There is no constraint on the number of links or the structure of the network that those links create (tree, directed acyclic graph, etc.) (0..*).
- hasContent: Relates a structural node to information units (e.g. a definition attached to a Terminological Entry). An instance of an information unit is attached to one and only one structural node (1..1).
- refinement: Relates information units that provide additional information to another information unit (IU) (e.g. a /note/ refining a /definition/). A refining IU is related to one and only one refined IU (1..1).

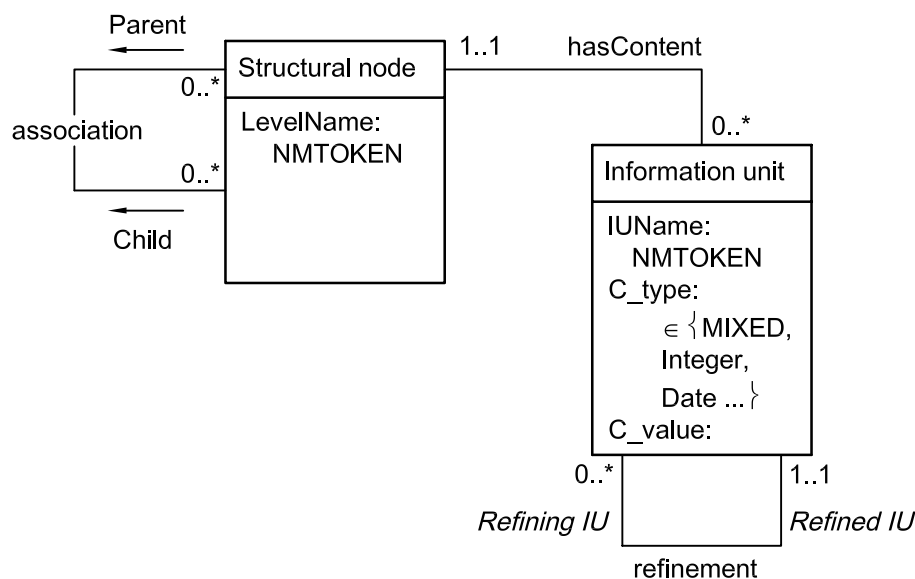


Figure 4 — UML diagram for structural nodes and information units

5.2.2 The MIXED type is an ordered combination of textual content and information units, corresponding to any kind of annotated content. It can be represented in a UML by means of the agglomeration operator, as shown in Figure 5.

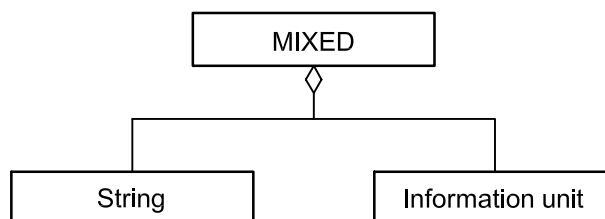


Figure 5 — MIXED object class

Keeping this rather general definition leads to the fact that annotations in textual content can be refined by other information units (for instance to indicate when and by whom the annotation has been made).

5.3 The terminological meta-model

The terminological meta-model is based on guidelines concerning the methods and principles of terminology management involving the production of terminological entries as described in ISO 704. One of the most important characteristics of a terminological entry — compared to a lexicographical entry — is its concept orientation. A terminological entry treats one concept in a given language and, in the case of multilingual

terminological entries, one or more totally or partially equivalent concepts in (an)other language(s), whereas a lexicographical entry contains one lemma (the base form of a single lexical unit) and one or more definitions (representing different meanings) in one or more languages. The meta-model presented in this International Standard gives guidelines for terminological entries. It is understood in this context that although lexicographical entries typically contain word or lexeme related information (such as part of speech, gender, etc.), some of this lexicographical information is frequently also included in terminological entries.

A terminological data collection comprises global information about the collection and a number of entries. Each entry performs three functions:

- it describes one concept, or two or more totally or partially equivalent concepts, in one or more languages;
- it lists the terms that designate the concept(s);
- it describes the terms themselves.

Each entry can have multiple language sections, and each language section can have multiple terminological units. Each data element in an entry can be associated with various kinds of descriptive and administrative information. In addition, there are various other resources that are not part of any one entry, but that can be linked to one or more entries. Such resources include bibliographic references, descriptions of ontologies, and binary data such as images that illustrate concepts.

By instantiating the generic architecture presented in 5.2, the terminological meta-model is described through seven instances from the structural node class.

- TDC (terminological data collection): Top level container for all information contained in a terminology system.
- GI (global information): Information that applies to all elements represented in a file, as opposed to information that may pertain to some but not to all components of the file. Usually contains, for example, the title of the (XML) file, the institution or individual originating the file, address information, copyright information, update information, etc.
- TE (terminological entry): Information that pertains to a single concept. Usually contains, for example, descriptive information pertinent to a concept, and administrative information concerning the concept. Can contain one or more language sections depending on whether the termbase is monolingual, bilingual, or multilingual.
- CI (complementary information): Usually contains, for example, textual bibliographical or administrative information residing in or external to the file, static or dynamic graphic images, video, audio, or virtually any other kind of binary data (i.e. blobs). Might also include references to other terminological resources or contextual links to related text corpora or to ontologies. These items are often designated as shared resources because they are available to all points in a termbase and are not repeated for different entries.
- LS (language section): Contains all the term sections for a terminological entry that are used in a given language, as well as information. Usually contains, for example, definitions, contexts, etc. associated with that language or the terms in that language.
- TS (term section): Information about terms. Usually contains, for example, a single term used to designate the concept that is the subject of the terminological entry, as well as any other information (e.g. definitions, contexts, etc.), associated with that term.
- TCS (term component section): Information about morphemic elements, words, or contiguous strings from which a polymorphemic (or multiword) term is formed. In some languages, such as German or English, it is frequently unnecessary to distinguish information about the individual components making up a polynomial term. In other languages, such as French or Spanish, it is important to be able to include information such as gender for the individual words used in constructing a multiword term because this information is necessary when using the term in texts.

These instances of structural levels implement the “association” relation with constraints on cardinality (see Figure 6), which can also be schematized by the sketch shown in Figure 7.

- A TE can contain any number of LSs (0..*).
- An LS can contain any number of TSs (0..*).