INTERNATIONAL STANDARD

ISO 24616

First edition
2012-09-01

Language resources management — Multilingual information framework

Gestion des ressources langagières — Plateforme d'informations multilingues

Reference number
ISO 24616:2012(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 24616 was prepared by Technical Committee ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 4, *Language resource management*.

Language resources management — Multilingual information framework

1 Scope

This International Standard provides a generic platform for modelling and managing multilingual information in various domains: localization, translation, multimedia annotation, document management, digital library support, and information or business modelling applications. MLIF (multilingual information framework) provides a metamodel and a set of generic data categories [ISO 12620:2009] for various application domains. MLIF also provides strategies for the interoperability and/or linking of models including, but not limited to, XLIFF, TMX, smilText and ITS.

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 12620:2009; *Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources*

ISO 8879, *Information processing — Text and office systems — Generalized Markup Language (SGML)*

Extensible Markup Language. Fifth Edition, T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, F. Yergeau Editors, W3C Recommendation, 26 November 2008, http://www.w3.org/TR/xml

3 Terms and definitions

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

3.1
adornment
data category attached to a component of a metamodel

3.2
inline code
inline instructions inserted in a source document

Note to entry: Native code can, for instance, provide presentational instructions (e.g. HTML codes).

3.3
subtitle
textual versions of the dialog in films, television programs, video games, etc., usually displayed at the bottom of the screen

3.4
working language
language in which linguistic sequences are expressed

4 Specification principles

4.1 Key standard used in the specification: Unified Modeling Language (UML)

The MLIF specification complies with the modelling principles of UML as defined by the Object Management Group (OMG) [UML]. The specification uses the UML subset that is relevant for the purposes of MLIF.

4.2 Metamodel and adornment

In line with Terminological Markup Framework (TMF) as defined in ISO 16642, MLIF defines a metamodel that is adorned by data categories, as defined in ISO 12620.

4.3 XML serialization

Associated with the metamodel and its adornment, MLIF proposes a representation in XML called “XML serialization”, in line with Extensible Markup Language (XML) as defined in ISO 8879.

5 Metamodel specification

The MLIF metamodel is specified in the UML object diagram in Figure 1.

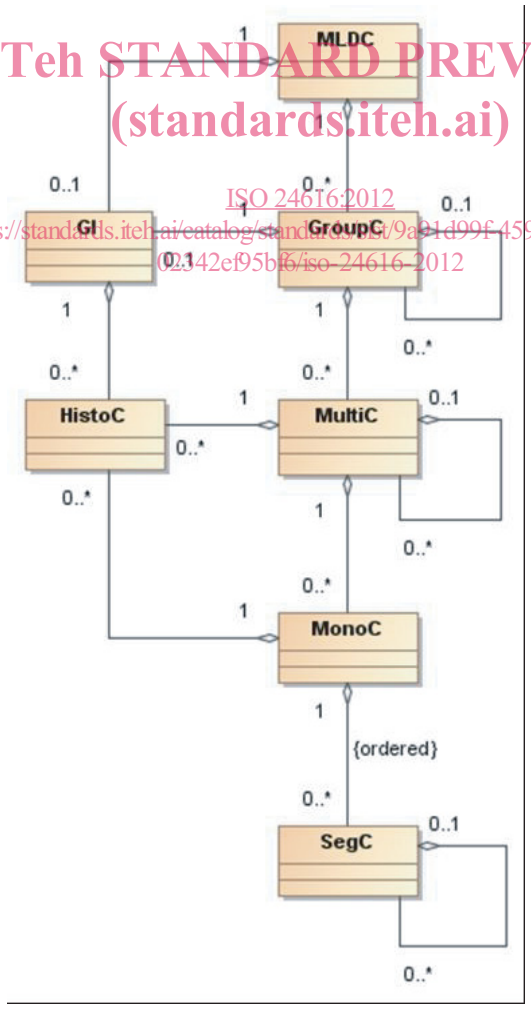

Figure 1 — MLIF metamodel

The MLIF metamodel is defined by the following seven "core components". These components are listed as follows, according to their XML serialization:

— <MLDC> (Multilingual Data Collection), which represents a collection of data containing global information and several multilingual units;

— <GI> (Global Information), which represents technical and administrative information applying to the entire multilingual data collection;

— <GroupC> (Grouping components), which represents a sub-collection of multilingual data that have a common origin or purpose within a given project;

— <MultiC> (Multilingual Component), which groups together all variants of a given textual content;

— <MonoC> (Monolingual Component), which groups together information related to one language and is part of a multilingual component (MultiC);

— <HistoC> (History Component), which traces modifications to the component to which it is anchored (i.e. versioning);

— <SegC> (Segmentation Component), which allows any level of segmentation for textual information, possibly in a recursive manner.

6 MLIF compliance

Any format compliant with this International Standard may use the MLIF metamodel in two possible ways:

— by fully implementing the MLIF metamodel starting at the level of <MLDC>;

— by specifically embedding MLIF-compliant information within another model, by implementing one of the lower level MLIF elements, namely <GroupC>, <MultiC> or <MonoC>.

7 Metamodel adornment

7.1 Introduction

The MLIF XML serialization proposes a set of XML elements and XML attributes, which are described in the following sections, where the characters “<” and “>” delimit the name of the element. Following the TEI guidelines (http://www.tei-c.org), some attributes are specified by means of a class attribute, with the convention that the name of the class attribute is prefixed by “att.” (e.g. “att.xlink”). The other XML attributes are listed with the convention that two quotes delimit the name of the attribute (e.g. “xml:lang”). The specifications in Annex G shall be applied.

7.2 General principles concerning the use of W3C generic attributes

The following W3C attributes are to be used by all MLIF-compliant applications:

— the attribute xml:lang shall be used in accordance with W3C recommendations to represent the working language of any relevant element and, in particular, shall be used systematically for any implementation of MonoC;

— the attribute xml:id shall be used in accordance with W3C recommendations to provide a unique identifier to an element of the MLIF metamodel.

7.3 Recommended adornment for GI

— <domain>

— <project>

— <source>

— <sourceType>

— <sourceLanguage>

— <sourceFormat>

— <targetLanguage>

— <formatVersion>

— <legalStatus>

— <creationTool>

— <creationToolVersion>

— <creationDate>

— <creationIdentifier>

— <changeDate>

— <changeIdentifier>

7.4 Recommended adornment for GroupC

— <groupType>

7.5 Recommended adornment for MultiC

— <class>

— <changeDate>

— <changeIdentifier>

— <creationTool>

— <creationToolVersion>

— <creationIdentifier>

— <creationDate>

— <translationStatus>

— <matchQuality>

7.6 Recommended and mandatory adornment for MonoC

— att.lang

— <translationRole>

— <segmentation>

— att.xlink

The language attribute is mandatory on MonoC. All other adornments are optional.

7.7 Recommended adornment for SegC

— <translationRole>

— <beginPairedTag>

— <endPairedTag>

— <genericGroupPlaceholder>

— <placeholder>

— <genericPlaceholder>

— <translate>

— att.linguistic

— att.xlink

7.8 Recommended adornment for HistoC

The HistoC component is a generic component that traces modifications made on the component to which it is anchored (e.g. creation, modification and validation). In the MLIF metamodel, the HistoC component may be anchored to the GI, MultiC or MonoC component. This makes it possible for all evolutions of, or enhancements to, the component to be recorded.

HistoC may be adorned by four elements:

— <author>

— <version>

— <transaction>

— <date>

7.9 Recommended online annotation adornment

Multilingual text documents are often only one stage in a complex workflow that involves external document sources in a wide variety of formats. From these, it is often necessary to keep inline markup indicating the presentational features that have to be retained in a translated target document. To this end, MLIF-compliant applications should use the following elements, in relation to the <SegC> element, that map onto similar subsets in TMX and XLIFF:

— <beginPairedTag>

— <endPairedTag>

— <genericGroupPlaceholder>

— <genericPlaceholder>

— <placeholder>

7.10 Recommended adornment for localization

All the following elements should be used to provide localization-related information:

— <translationRole>

— <translationStatus>

7.11 Recommended adornment for internationalization

— <translate>

7.12 Recommended adornment for temporal synchronization

The following elements should be used when textual content has to be conveyed (in written or spoken form) together with some constraints:

— <duration>

— <begin>

— <next>

8 Relation with other standards

As with the “Terminological Markup Framework” TMF [ISO 16642] in terminology, MLIF introduces a metamodel that combines with selected data categories as a way of ensuring interoperability between several multilingual applications and corpora. MLIF deals with multilingual corpora, multilingual fragments, and the translation relations between them. In each domain where MLIF is applicable, a specific granularity may be considered for segmentation and description. These two last processes may rely on MAF [ISO 24611], SynAF [ISO 24615] and TMF for morphological description, syntactical annotation and terminological description respectively.

MLIF supports the construction and the interoperability of localization and translation memories resources, and also deals with the description of a metamodel for multilingual content. MLIF does not propose a closed list of description features. Rather, it provides a list of data categories that is much easier to update and extend. This list represents a point of reference for multilingual information in the context of various application scenarios.

However, MLIF not only describes elementary linguistic segments (e.g. sentence, syntactical fragment, word and part of speech), but may also be used to represent document structure (e.g. title, abstract, paragraph and section). In addition, MLIF allows for external and internal links (annotations and references).

MLIF is designed to provide a common framework that facilitates the interoperability with formats such as TMX (LISA OSCAR) and XLIFF (OASIS). MLIF can be seen as a parent of these formats, since both of them

deal with multilingual data expressed in the form of segments or text units. Both can be stored, manipulated and translated in a similar manner.

Examples of using MLIF are given in Annexes A to F.

Annex A
(informative)

Example using MLIF for Computer-Assisted Translation (CAT)

The main reason for lemma, part-of-speech and morphological features is to allow CAT tools based on translation memory to produce translations of new words and sentences that are not in the translation database.

For example, using a translation memory that contains the English sentence "The meal is nice." and its translation in French "Le repas est bon.", current CAT tools such as SDL TRADOS[1)] Translator's Workbench are not able to provide the predicted translation for the sentence "The meals are nice." even though the word lemmas of "The meal is nice." and "The meals are nice." are matching. This weakness is due to the fact that these tools use limited linguistic criteria during the translation process.

The data produced by TRADOS Translator's Workbench is as follows:

```
<tmx version="1.4">
 <header
  creationtool="TRADOS Translator's Workbench for Windows"
  creationtoolversion="Edition 8 Build 863"
  segtype="sentence"
  o-tmf="TW4Win 2.0 Format"
  adminlang="EN-US"
  srclang="EN-GB"
  datatype="rtf"
  creationdate="20100528T144322Z"
  creationid="USER"/>
 <body>
  <tu creationdate="20100528T144322Z" creationid="USER">
   <tuv xml:lang="EN-GB">
    <seg>The meal is nice.</seg>
   </tuv>
   <tuv xml:lang="FR-FR">
    <seg>Le repas est bon.</seg>
   </tuv>
  </tu>
 </body>
</tmx>
```


To translate the sentence "The meals are nice.", an MLIF-compliant tool should implement the following procedure:

Step-1 Represent in MLIF and add linguistic properties to all the words within the translation memory.

Step-2 Run a part-of-speech tagger on the sentence in order to obtain the right morphosyntactic word categories.

Step-3 Translate the lemmas using an English-to-French bilingual lexicon.

1) SDL TRADOS Translator's Workbench is an example of a suitable product available commercially. This information is given for the convenience of users of this International Standard and does not constitute an endorsement by ISO of this product.

Step-4 Consult a French lexicon of inflected forms in order to retrieve the correct inflected form using the lemma and morphological features.

Step-5 Generate the translation of "The meals are nice." by substituting each English word with its French inflected form as follows:

"The meals are nice." => "Les repas sont bons."

The XML data will include a feature structure declaration defining a tagset (e.g. for "nS"), with a word segmentation and tagset defined in MAF:

```
<MLDC xmlns="http://www.tei-c.org/ns/1.0">
 <tei:fLib>
  <tei:f xml:id="nS" name="grammaticalNumber" fVal="singular"/>
  <tei:f xml:id="gM" name="grammaticalGender" fVal="masculine"/>
  <tei:f xml:id="mP" name="verbFormMood" fVal="present"/>
  <tei:f xml:id="p1" name="person" fVal="thirdPerson"/>
  <tei:f xml:id="nS" name="grammaticalNumber" fVal="singular"/>
 </fLib>
 <GroupC>
  <MultiC>
   <creationIdentifier>SEMMAR</creationIdentifier>
   <creationDate>20090922T140653Z</creationDate>
   <MonoC xml:lang="en">
    <SegC>The meal is nice.</SegC>
   </MonoC>
   <MonoC xml:lang="fr">
    <SegC>Le repas est bon.</SegC>
   </MonoC>
  </MultiC>
  <MultiC class="translation">
   <MonoC xml:lang="en">
    <SegC class="word" lemma="the" pos="definiteArticle">The</SegC>
    <SegC
     class="word"
     lemma="meal"
     pos="commonNoun"
     tag="#nS">meal</SegC>
    <SegC
     class="word"
     lemma="be"
     pos="verb"
     tag="#mP #p1 #nS">is</SegC>
    <SegC class="word" lemma="nice" pos="qualifierAdjective">nice</SegC>
    <SegC class="word" lemma="." pos="mainPunctuation">.</SegC>
   </MonoC>
   <MonoC xml:lang="fr">
    <SegC
     class="word"
     lemma="le"
     pos="definiteArticle"
     tag="#gM #nS">Le</SegC>
    <SegC
     class="word"
     lemma="repas"
     pos="commonNoun"
     tag="#gM #nS">repas</SegC>
    <SegC
     class="word"
     lemma="être"
```