МЕЖДУНАРОДНЫЙ СТАНДАРТ

ISO 24616

Первое издание
2012-09-01

Управление языковыми ресурсами. Многоязыковая информационная система

Language resources management – Multilingual information framework

Ответственность за подготовку русской версии несёт GOST R (Российская Федерация) в соответствии со статьёй 18.1 Устава ISO

Ссылочный номер
ISO 24616:2012(R)

Бюро ISO по авторским правам:
Case postale 56 • CH-1211 Geneva 20
Тел.: + 41 22 749 01 11
Факс: + 41 22 749 09 47
Эл. почта: copyright@iso.org
Веб-сайт: www.iso.org

Опубликовано в Швейцарии

Содержание

Страница

Предисловие

Международная организация по стандартизации (ISO) является всемирной федерацией национальных организаций по стандартизации (комитетов-членов ISO). Разработка международных стандартов обычно осуществляется техническими комитетами ISO. Каждый комитет-член, заинтересованный в деятельности, для которой был создан технический комитет, имеет право быть представленным в этом комитете. Международные правительственные и неправительственные организации, имеющие связь с ISO, также принимают участие в работе. ISO работает в тесном сотрудничестве с Международной электротехнической комиссией (IEC) по всем вопросам стандартизации в области электротехники.

Проекты международных стандартов разрабатываются согласно правилам, приведённым в Директивах ISO/IEC, Часть 2.

Разработка международных стандартов является основной задачей технических комитетов. Проекты международных стандартов, принятые техническими комитетами, рассылаются комитетам-членам на голосование. Для публикации в качестве международного стандарта требуется одобрение не менее 75 % комитетов-членов, принявших участие в голосовании.

Принимается во внимание тот факт, что некоторые из элементов настоящего документа могут быть объектом патентных прав. ISO не принимает на себя обязательств по определению отдельных или всех таких патентных прав.

ISO 24616 был подготовлен Техническим комитетом ISO/TC 37, *Терминология и другие языковые и информационные ресурсы,* Подкомитетом SC 4, *Управление языковыми ресурсами*.

Управление языковыми ресурсами. Многоязыковая информационная система

1 Область применения

Настоящий Международный стандарт обеспечивает универсальную платформу для моделирования многоязыковой информации и управления ею в самых разных сферах: локализации, перевода, мультимедийного аннотирования, организации документооборота, ведения цифровых библиотек и в прикладных системах моделирования хозяйственной деятельности предприятий. Многоязыковая информационная система MLIF (multilingual information framework) предоставляет соответствующую высокоуровневую модель (метамодель) и множество универсальных категорий данных [согласно ISO 12620:2009] для многочисленных прикладных областей. Она обеспечивает также необходимые стратегии взаимодействия и/или связывания различных моделей, включая, в частности, широко используемые модели XLIFF, TMX, smilText и ITS.

2 Нормативные ссылки

Перечисленные ниже ссылочные документы обязательны для применения данного документа. В случае датированных ссылок действующим является только указанное издание. Применительно к недатированным ссылочным документам применяются их самые последние издания (включая все последующие изменения):

ISO 12620:2009; *Терминология, другие языковые ресурсы и ресурсы содержания. Спецификация категорий данных и ведение реестра категорий данных для языковых ресурсов*

ISO 8879, *Обработка информации. Текстовые и офисные системы. Стандартный обобщённый язык разметки (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 Термины и определения

В рамках настоящего документа используются термины и определения, приведённые ниже:

3.1
стилистический орнамент
adornment
категория данных, приписываемая компоненту метамодели

3.2
внутритекстовый код
inline code
внутритекстовые команды, встроенные в исходный документ

Примечание к статье: на естественном языке могут записываться, в частности, команды представления информации (например, коды HTML).

3.3
субтитр
subtitle
текстовые эквиваленты диалогов в кинофильмах, телепрограммах, видеоиграх и т.п., обычно отображаемые внизу экрана

3.4
рабочий язык
working language
язык, с помощью которого выражаются последовательности лингвистических единиц

4 Принципы описания

4.1 Основополагающий стандарт спецификаций: универсальный язык моделирования UML

В основе спецификации MLIF лежат принципы построения моделей на языке UML, как он был определён Группой объектного управления OMG [Object Management Group]. В спецификации используется подмножество элементов языка UML, подходящее для целей MLIF.

4.2 Метамодель и стилистический орнамент

Наряду с терминологической системой разметки TMF (Terminological Markup Framework), как она определена в ISO 16642, MLIF определяет метамодель, орнаментированную категориями данных, как она представлена в ISO 12620.

4.3 XML-сериализация

Совместно с метамоделью и её стилистическим орнаментом MLIF даёт представление информации на языке XML, называемое “XML-сериализацией”, в сочетании с расширяемым языком разметки XML (Extensible Markup Language), как он определён в ISO 8879.

5 Спецификация метамодели

Метамодель MLIF описывается объектной диаграммой на языке UML, показанной на Рисунке 1. Эту модель определяют следующие семь “компонентов ядра”, перечисленных ниже в порядке их XML-сериализации:

— <MLDC> (Multilingual Data Collection / Многоязыковая коллекция данных), которая представляет собой совокупность данных, содержащих информацию глобального характера и несколько многоязыковых лингвистических единиц;

— <GI> (Global Information / Глобальная информация), включающая в себя сведения технического и административного характера, применимые ко всей коллекции многоязыковых данных в целом;

— <GroupC> (Grouping components / Компоненты группирования), которые представляют подмножество многоязыковых данных, имеющих общий источник или общее целевое назначение в рамках конкретного проекта;

— <MultiC> (Multilingual Component / Многоязыковой компонент), обеспечивающий группировку всех вариантов определённого текстового содержания;

— <MonoC>(Monolingual Component / Одноязычный компонент), обеспечивающий группировку информации, которая относится к одному и тому же языку и является частью многоязыкового компонента MultiC;

— <HistoC> (History Component / Компонент предыстории), обеспечивающий отслеживание изменений компонента, к которому он привязан (т.е. контроль версий);

— <SegC> (Segmentation Component / Компонент сегментации), который обеспечивает возможность любого уровня сегментирования текстовой информации, в том числе – с использованием рекурсивного метода.

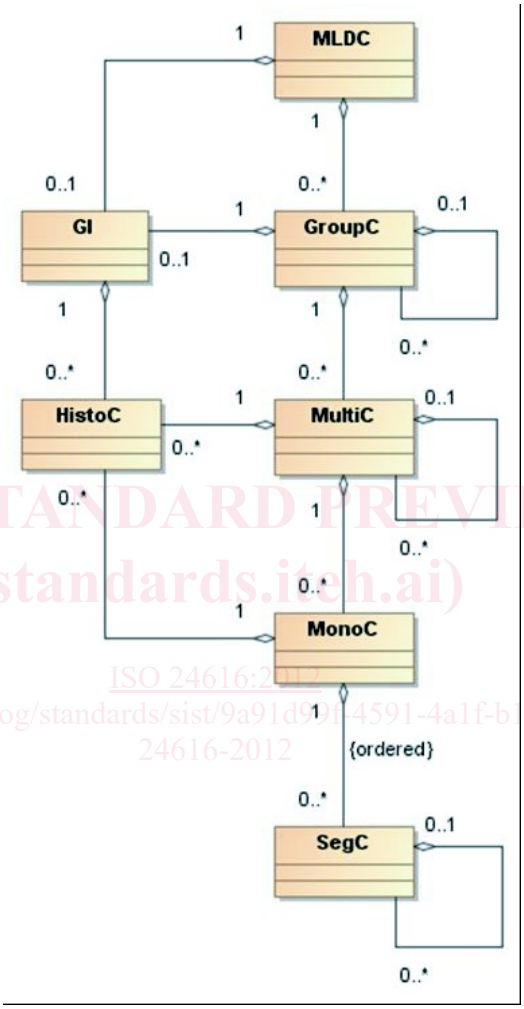

Рисунок 1 — Схематическое представление метамодели MLIF

6 Применимость MLIF

Метамодель MLIF может использоваться применительно к любому формату, совместимому с настоящим международным стандартом, двумя способами:

— посредством полной реализации метамодели MLIF, начиная с уровня <MLDC>;

— путём специального вложения информации, совместимой с MLIF, в другую модель, с целью реализации низкоуровневых элементов MLIF, а именно <GroupC>, <MultiC> или <MonoC>.

7 Стилистический орнамент метамодели

7.1 Общие замечания

XML-сериализация MLIF предполагает наличие множества элементов и атрибутов XML, которые описываются в последующих разделах настоящего стандарта и в которых символы “<” и “>” ограничивают имя элемента. В соответствии с руководящими указаниями TEI (http://www.tei-c.org), некоторые из атрибутов определяются путём указания их класса, и в этом случае атрибут имени класса предваряется префиксом “att.” (например “att.xlink”). В то же время другие атрибуты XML определяются списком, в котором имена атрибутов выделяются кавычками (например “xml:lang”). При этом должны использоваться спецификации, представленные в Приложении G настоящего стандарта.

7.2 Общие принципы использования групповых атрибутов W3C

Во всех MLIF-совместимых приложениях подлежат использованию следующие атрибуты, определённые консорциумом W3C:

— атрибут xml:lang должен применяться для представления рабочего языка любого релевантного элемента и, в частности, использоваться систематически при любой реализации компонента MonoC;

— атрибут xml:id должен использоваться в соответствии с рекомендациями W3C для предоставления уникального идентификатора элемента метамодели MLIF.

7.3 Рекомендуемый стилистический орнамент для компонента GI

— <domain>

— <project>

— <source>

— <sourceType>

— <sourceLanguage>

— <sourceFormat>

— <targetLanguage>

— <formatVersion>

— <legalStatus>

— <creationTool>

— <creationToolVersion>

— <creationDate>

— <creationIdentifier>

— <changeDate>

— <changeIdentifier>

7.4 Рекомендуемый стилистический орнамент для компонента GroupC

— <groupType>

7.5 Рекомендуемый стилистический орнамент для компонента MultiC

— <class>

— <changeDate>

— <changeIdentifier>

— <creationTool>

— <creationToolVersion>

— <creationIdentifier>

— <creationDate>

— <translationStatus>

— <matchQuality>

7.6 Рекомендуемый стилистический орнамент для компонента MonoC

— att.lang

— <translationRole>

— <segmentation>

— att.xlink

Для компонента MonoC обязательно наличие языкового атрибута; все другие атрибуты - факультативны.

7.7 Рекомендуемый стилистический орнамент для компонента SegC

— <translationRole>

— <beginPairedTag>

— <endPairedTag>

— <genericGroupPlaceholder>

— <placeholder>

— <genericPlaceholder>

— <translate>

— att.linguistic

— att.xlink

7.8 Рекомендуемый стилистический орнамент для компонента HistoC

Групповой компонент HistoC обеспечивает отслеживание изменений компонента, к которому он привязан (например, этапов создания, модификации и подтверждение достоверности). В метамодели MLIF компонент HistoC может привязываться к компоненту GI, MultiC или MonoC, благодаря чему становится возможной регистрация всех эволюционных изменений или расширений компонента.

Компонент HistoC может иметь стилистический орнамент из четырёх элементов:

— <author>

— <version>

— <transaction>

— <date>

7.9 Рекомендуемый стилистический орнамент для оперативно доступной аннотации

Многоязычные текстовые документы часто появляются только на одном этапе сложного технологического процесса, в котором участвуют внешние источники документов, имеющих самые разные форматы. Отсюда часто возникает необходимость сохранения внутритекстовой разметки, указывающей на характеристики представления данных, которые подлежат сохранению и в целевом документе на языке перевода. Поэтому в рамках MLIF-совместимых приложений применительно к компоненту <SegC> должны использоваться следующие элементы, отображаемые на аналогичные подмножества элементов в TMX и XLIFF:

— <beginPairedTag>

— <endPairedTag>

— <genericGroupPlaceholder>

— <genericPlaceholder>

— <placeholder>

7.10 Рекомендуемый стилистический орнамент для локализации

Для предоставления необходимой информации, касающейся локализации, подлежат использованию следующие элементы:

— <translationRole>

— <translationStatus>

7.11 Рекомендуемый стилистический орнамент для интернационализации

— <translate>

7.12 Рекомендуемый стилистический орнамент для синхронизации во времени

Когда текстовое содержание документа подлежит передаче (в письменной или устной форме) вместе с некоторыми сопутствующими ограничениями, должны использоваться элементы:

— <duration>

— <begin>

— <next>

8 Связь с другими стандартами

Применительно к структуре терминологической разметки TMF [ISO 16642] при работе с терминологией MLIF предоставляет метамодель, которая в сочетании с выбранными категориями данных образует надёжную основу для обеспечения надлежащего взаимодействия между несколькими многоязыковыми приложениями в рамках работы с текстовыми корпусами. При этом MLIF обеспечивает работу с многоязыковыми корпусами, многоязычными фрагментами и отношениями, характеризующими перевод с одного языка на другой. В любой сфере применимости MLIF для целей сегментирования и описания текстов может выбираться определённый уровень разбиения текстовой информации. В этой части процессы сегментирования и описания могут основываться на использовании MAF [ISO 24611], SynAF [ISO 24615] и структуры терминологической разметки (TMF) для морфологического описания, синтаксического аннотирования и терминологического описания, соответственно.

MLIF поддерживает процессы разработки и взаимодействия ресурсов памяти переводов и процедур локализации, а также работу с описанием метамодели в части обработки её многоязычного контента. MLIF не предоставляет исчерпывающего списка характеристик используемых описаний, а вместо этого даёт перечень категорий данных, который гораздо более удобен для обновления и расширения. Этот перечень является отправной точкой для обработки многоязычной информации в контексте различных сценариев, реализуемых приложениями.

Однако MLIF не только описывает элементарные лингвистические сегменты (например, предложение, синтаксический фрагмент, слово и часть речи), но может также использоваться для представления структуры документа (например, заголовка, резюме, абзаца и раздела). Кроме того, MLIF допускает установление внешних и внутренних связей (аннотаций и ссылок).

MLIF предназначается для создания общей основы, облегчающей работу с такими форматами, как TMX (LISA OSCAR) и XLIFF (OASIS). MLIF может рассматриваться как родительский узел этих форматов, поскольку оба они относятся к многоязычным данным, выраженным в форме сегментов или текстовых единиц. Оба этих формата могут храниться, использоваться и преобразовываться одинаковым образом.

Примеры использования MLIF приведены в Приложениях от A до F.

Приложение A
(информативное)

Пример использования MLIF для автоматизированного перевода

Главная цель использования таких структур, как лемма, часть речи и морфологические элементы состоит в том, чтобы придать инструментальным средствам автоматизации перевода (CAT), основой которых является память переводов, способность к выполнению перевода новых слов и предложений, которые не содержатся в базе данных автоматизированной системы перевода.

Например, такая современная система памяти переводов, как SDL TRADOS [1)], в которой будут записаны английское предложение "The meal is nice" (“Эта еда великолепна”) и его перевод на французский язык "Le repas est bon", не способна будет дать очевидный перевод предложения "The meals are nice", несмотря на то, что текстовые леммы "The meal is nice" и "The meals are nice" полностью совпадают. Причина такой слабости кроется в том факте, что в данной системе автоматизации в процессе перевода задействовано недостаточное число лингвистических критериев.

В рассматриваемом случае данные, формируемые модулем TRADOS Translator's Workbench, выглядят следующим образом:

```
<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>
```


Для обеспечения перевода предложения "The meals are nice", MLIF-совместимое инструментальное средство должно было бы реализовать следующую процедуру:

Шаг 1 Представить в рамках MLIF с добавлением лингвистических характеристик все слова, хранящиеся в памяти переводов.

1) Система SDL TRADOS Translator's Workbench взята как подходящий для примера коммерческий программный продукт, широко доступный для приобретения. Информация приведена здесь для удобства пользователей настоящего Международного стандарта и не должна рассматриваться как одобрение указанной системы со стороны ISO.

Шаг 2 Пропустить предложение через программу частеречной разметки для получения правильных морфосинтаксических категорий слов.

Шаг 3 Осуществить перевод лемм с использованием двуязычного англо-французского словаря.

Шаг 4 Обратиться к французскому словарю форм склонения для выбора правильной словоформы по заданной лемме и морфологическим признакам.

Шаг 5 Сформировать переводной эквивалент фразы "The meals are nice" путём замены каждого английского слова его французской формой склонения следующим образом:

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

Данные на языке XML должны включать в себя объявление признаковой структуры, определяющее набор тегов (например для "nS"), и сегментацию слов с использованием набора тегов, определённого в рамках 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"
```