INTERNATIONAL STANDARD

ISO 24613

First edition
2008-11-15

Language resource management — Lexical markup framework (LMF)

Gestion de ressources langagières — Cadre de balisage lexical (LMF)

Reference number
ISO 24613:2008(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

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 24613 was prepared by Technical Committee ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 4, *Language resource management*.

ISO 24613 is designed to coordinate closely with ISO 12620, *Terminology and other content and language resources — Data categories — Specification of data categories and management of a Data Category Registry for language resources*[1)], and ISO 16642, *Computer applications in terminology — Terminological markup framework*.

1) To be published. (Revision of ISO 12620:1999)

Introduction

Optimizing the production, maintenance and extension of electronic lexical resources is one of the crucial aspects impacting human language technologies (HLT) in general and natural language processing (NLP) in particular, as well as human-oriented translation technologies. A second crucial aspect involves optimizing the process leading to their integration in applications. Lexical Markup Framework (LMF) is an abstract metamodel that provides a common, standardized framework for the construction of computational lexicons. LMF ensures the encoding of linguistic information in a way that enables reusability in different applications and for different tasks. LMF provides a common, shared representation of lexical objects, including morphological, syntactic and semantic aspects.

The goals of LMF are to provide a common model for the creation and use of electronic lexical resources ranging from small to large in scale, to manage the exchange of data between and among these resources, and to facilitate the merging of large numbers of different individual electronic resources to form extensive global electronic resources. The ultimate goal of LMF is to create a modular structure that will facilitate true content interoperability across all aspects of electronic lexical resources.

The LMF core package describes the basic hierarchy of information of a lexical entry, including information on the form. The core package is supplemented by various resources that are part of the definition of LMF. These resources include:

— specific data categories used by the variety of resource types associated with LMF, both those data categories relevant to the metamodel itself, and those associated with the extensions to the core package;

— the constraints governing the relationship of these data categories to the metamodel and to its extensions;

— standard procedures for expressing these categories and thus for anchoring them on the structural skeleton of LMF and relating them to the respective extension models;

— the vocabularies used by LMF to express related informational objects for describing how to extend LMF through linkage to a variety of specific resources (extensions) and methods for analysing and designing such linked systems.

Extensions of the core package which are documented in the annexes of this International Standard include:

a) machine readable dictionaries;

b) natural language processing lexical resources.

LMF extensions are expressed in a framework that describes the reuse of the LMF core components (such as structures, data categories, and vocabularies) in conjunction with the additional components required for a specific resource.

Types of individual instantiations of LMF can include such electronic lexical resources as fairly simple lexical databases, NLP and machine-translation lexicons, as well as electronic monolingual, bilingual and multilingual lexical databases. LMF provides general structures and mechanisms for analysing and designing new electronic lexical resources, but LMF does not specify the structures, data constraints and vocabularies to be used in the design of specific electronic lexical resources. LMF also provides mechanisms for analysing and describing existing resources using a common descriptive framework. For the purpose of both designing new lexical resources and describing existing lexical resources, LMF defines the conditions that allow the data expressed in any one lexical resource to be mapped to the LMF framework, and thus provides an intermediate format for lexical data exchange.

Language resource management — Lexical markup framework (LMF)

1 Scope

This International Standard describes the Lexical Markup Framework (LMF), a metamodel for representing data in lexical databases used with monolingual and multilingual computer applications.

LMF provides mechanisms that allow the development and integration of a variety of electronic lexical resource types[2]. These mechanisms will present existing lexicons as far as possible. If this is impossible, problematic information will be identified and isolated.

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 (all parts), *Codes for the representation of names of languages*

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

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

ISO 12620, *Terminology and other content and language resources — Data categories — Specification of data categories and management of a Data Category Registry for language resources* [3]

ISO 15924, *Information and documentation — Code for the representation of names of scripts*

3 Terms and definitions

For the purposes of this International Standard, the terms and definitions given in ISO 1087-1, ISO 1087-2 and the following apply [4].

3.1
abbreviated form
form (3.14) resulting from the omission of any part of the **full form** (3.16) of the same **lexeme** (3.25)

---

2) LMF supports existing lexical resource models such as the Genelex [9], the EAGLES International Standards for Language Engineering (ISLE) [5] and Multilingual ISLE Lexical Entry (MILE) models [6].

3) To be published. (Revision of ISO 12620:1999)

4) It is worth noting that we have purposely avoided defining and using highly controversial terms such as “word”, “morpheme”, “base”, “fusion”, “ergative”, “paradigm”, and “collocation”.

3.2
adjunct
non-essential element associated with a verb as opposed to **syntactic arguments** (3.43)

EXAMPLE Alfred (syntactic argument) reads a book (syntactic argument) today (adjunct).

NOTE Adverbs are possible adjuncts for a sentence.

3.3
affix
bound morph (3.8) that may contribute to a **form** (3.14) and participates in the process of **inflection** (3.20), **agglutination** (3.5), **derivation** (3.12) or **composition** (3.9)

NOTE Affixes function as prefixes (pre-positioned), suffixes (post-positioned), infixes (inserted) and circumfixes (combination of prefix and suffix).

3.4
affixation
process in which an **affix** (3.3) is added to a **lemma** (3.24) or a **stem** (3.40)

3.5
agglutination
process in which an **agglutinated form** (3.6) is made up

3.6
agglutinated form
word form (3.47) that a **lexeme** (3.25) can take when used in a sentence or a phrase within an **agglutinating language** (3.7)

3.7
agglutinating language
language where the different **word forms** (3.47) of the same **lexeme** (3.25) exhibit a variation and that may consist of more than one **morph** (3.31) but the boundaries between morphs are always clear-cut

EXAMPLE Korean, Japanese, Hungarian and Turkish are agglutinating languages [16].

3.8
bound morph
morph that appears only together with one or several other **morphs** (3.31)

3.9
composition
compounding
lexeme (3.25) formation in which a new lexeme [associated with its **part of speech** (3.37) information] is formed by adjoining at least two lexemes, in their original **forms** (3.14) or with slight transformations

NOTE Composition should not be confused with agglutination and derivation, where bound morphs are added to free ones.

3.10
compound
lexeme (3.25) associated with **part of speech** (3.37) information that is built from two or more lexemes

3.11
compound form
form (3.14) resulting from a **composition** (3.9)

3.12
derivation
change in the **forms** (3.14) of a **lexeme** (3.25) to create a new lexeme, usually by modifying the **stem** (3.40) or by **affixation** (3.4)

NOTE Sometimes derivation signals a change in part of speech, such as *nation* to *nationalize*. Sometimes the part of speech remains the same as in *nationalization* vs. *denationalization*.

3.13
derived form
form (3.14) resulting from a **derivation** (3.12)

3.14
form
sequence of **morphs** (3.31)

3.15
free morph
morph (3.31) that may stand by itself

EXAMPLE The English noun *boy.*

3.16
full form
complete representation of a **lexeme** (3.25) for which there is an **abbreviated form** (3.6)

3.17
grammatical feature
property associated to the **inflected** (3.19), **agglutinated** (3.6), **compound** (3.11) or **derived form** (3.13) that describes the grammatical attribute of the form

NOTE An example of a grammatical feature is: /grammatical gender/. (Following the convention adopted in the revision of ISO 12620, the slashes are used in order to delimit data category values.)

3.18
graph
minimal unit in a written language including letters, pictograms, ideograms, numerals and punctuations

3.19
inflected form
word form (3.47) that a **lexeme** (3.25) can take when used in a sentence or a phrase within an **inflectional language** (3.21)

3.20
inflection
process in which an **inflected form** (3.19) is made up

3.21
inflectional language
inflecting language
language where the different **word forms** (3.47) of the same **lexeme** (3.25) exhibit a variation and where there is no clear-cut boundary between **morphs** (3.31) in that morphs are generally fused together to yield a single, non-segmentable **form** (3.14)

EXAMPLE Spanish, Italian, French and English are inflectional languages [16].

3.22
interlingua
abstract intermediary language used in the machine translation of human languages

3.23
isolating language
language where the vast majority of **morphs** (3.31) are **free morphs** (3.15)

EXAMPLE Chinese is an isolating language.

3.24
lemma
lemmatized form
canonical form
conventional **form** (3.14) chosen to represent a **lexeme** (3.25)

EXAMPLE In European languages, the lemma is usually the /singular/ if there is a variation in /number/, the /masculine/ form if there is a variation in /gender/ and the /infinitive/ for all verbs. In some languages, certain nouns are defective in the singular form, in which case, the /plural/ is chosen. In Arabic, for a verb, the lemma is usually considered as being the third person singular with the accomplished aspect.

3.25
lexeme
abstract unit generally associated with a set of **forms** (3.14) sharing a common meaning

3.26
lexical entry
container for managing one or several **forms** (3.14) and possibly one or several meanings in order to describe a **lexeme** (3.25)

3.27
lexical resource
lexical database
database consisting of one or several **lexicons** (3.28)

3.28
lexicon
resource comprising **lexical entries** (3.26) for a given language

NOTE A special language lexicon or a lexicon prepared for a specific NLP application can comprise a specific subset of language.

3.29
machine readable dictionary
MRD
electronic **lexical resource** (3.27) designed to be consulted by human beings

NOTE Historically, MRDs were first computer representations of “printed” dictionaries, that's why they are called *machine readable* now.

3.30
machine translation lexicon
electronic **lexical resource** (3.27) in which the individual **lexical entries** (3.26) contain equivalents in two or more languages together with morphological, syntactic and/or semantic information to facilitate automatic or semi-automatic processing of **lexemes** (3.25) during machine translation

3.31
morph
sequence of **graphs** (3.18) or sequence of **phones** (3.38)

EXAMPLE The word *boys* consists of two morphs: *boy* and *s*.

3.32
morphological pattern
set of associations and/or operations that build the various forms of a **lexeme** (3.25), possibly by **inflection** (3.20), **agglutination** (3.5), **composition** (3.9) or **derivation** (3.12), depending on the language

NOTE A morphological pattern is not the explicit list of inflected forms. It usually references a prototypical class of inflectional forms, e.g. *ring,* as per *sing.*

3.33
morphology
description of the structure and formation of **forms** (3.14)

3.34
multiword expression
MWE
lexeme (3.25) made up of a sequence of two or more lexemes that has properties that are not predictable from the properties of the individual lexemes or their normal mode of combination

NOTE An MWE can be a compound, a fragment of a sentence, or a sentence. The group of lexemes making up an MWE can be continuous or discontinuous. It is not always possible to mark an MWE with a part of speech.

EXAMPLE "To kick the bucket", which means to die rather than to hit a bucket with one's foot.

3.35
natural language processing
NLP
field covering knowledge and techniques involved in the processing of linguistic data by a computer

3.36
orthography
way of spelling or writing **lexemes** (3.25) that conforms to a conventionalized use

NOTE Aside from standardized spellings of alphabetical languages, such as standard UK or US English, or reformed German spelling, there can be variations such as transliterations of languages in non-native scripts, stenographic renderings, or representations in the International Phonetic Alphabet. In this regard, orthographic information in a lexical entry can describe a kind of transformation applied to the form that is the object of the entry.

3.37
part of speech
lexical category
word class
category assigned to a **lexeme** (3.25) based on its grammatical properties

NOTE Typical parts of speech for European languages include: *noun, verb, adjective, adverb, preposition*, etc.

3.38
phone
minimal unit in the sound system of a language

3.39
script
set of graphic characters used for the written **form** (3.14) of one or more languages

[ISO/IEC 10646:2003, definition 4.37]

NOTE The description of scripts ranges from a high level classification such as hieroglyphic or syllabic writing systems vs. alphabets to a more precise classification like Roman vs. Cyrillic. Scripts are defined by a list of values taken from ISO 15924.

EXAMPLE Hiragana, Katakana, Latin and Cyrillic.

3.40
stem
sequence of **morphs** (3.31) that is smaller than or equal to the **form** (3.14) of a single **lexeme** (3.25) and that may be affected by an **inflectional** (3.20), **agglutinative** (3.5), **compositional** (3.9) or **derivation** (3.12) process

3.41
subcategorization frame
valence
valency
set of restrictions on a **lexeme** (3.25) indicating the properties of the **syntactic arguments** (3.43) that can or must occur with this given lexeme

3.42
support verb
verb that makes a generic semantic contribution to the context and that combines with a noun to form a **lexeme** (3.25)

EXAMPLE *take an exam* or *give an exam*. In these examples, *take* and *give* have only limited inherent meaning based on their semantics, but rather are used in a conventional, generic way to express a collocational conceptualization.

3.43
syntactic argument
one of the essential and functional elements in a clause that identifies the participants in the process referred to by a verb

EXAMPLE Alfred (syntactic argument) reads a book (syntactic argument) today (adjunct).

3.44
transcription
form (3.14) resulting from a coherent method of writing down speech sounds, to include converting speech sounds described in one writing system to an equivalent representation of the same speech sounds described in another writing system

3.45
transliteration
form (3.14) resulting from the conversion of one writing system into another, usually through a one to one correspondence between characters

3.46
variant
one of the alternative **forms** (3.14) of a **lexeme** (3.25)

3.47
word form
form (3.14) that a **lexeme** (3.25) takes when used in a sentence or a phrase

4 Key standards used by LMF

4.1 Unicode

LMF is Unicode compliant and presumes that all data are represented using Unicode character encodings.

4.2 Language coding

Language identifiers used in LMF-compliant resources shall conform to criteria specified in the ISO 639 family of standards. Some issues involving the combination of language and country codes, as well as the coordination of different parts of ISO 639 have been addressed in external standards supported by the

technology community. It is recommended that users consult the current edition of IETF Best Common Practices (BCP) 47, *Tags for the Identification of Languages* in order to resolve issues involving choosing and matching identifiers for use in electronic environments [1].

4.3 Script Coding

When the script code is not part of the language identifier, script identifiers shall conform to criteria specified in ISO 15924.

4.4 ISO 12620 Data Category Registry (DCR)

The designers of an LMF conformant lexicon shall use data categories from the ISO 12620 Data Category Registry (DCR) located at www.isocat.org.

4.5 Unified Modeling Language (UML)

LMF complies with the specifications and modeling principles of UML as defined by the Object Management Group (OMG) [2]. LMF uses a subset of UML that is relevant for linguistic description.

5 The LMF model

5.1 Introduction

LMF models are represented by UML classes, associations among the classes, and a set of ISO 12620 data categories that function as UML attribute-value pairs. The data categories are used to adorn the UML diagrams that provide a high level view of the model. LMF specifications in the form of textual descriptions that describe the semantics of the modeling elements provide more complete information about classes, relationships, and extensions than can be included in UML diagrams.

In this process, lexicon developers shall use the classes that are specified in the **LMF core package** (5.2). Additionally, developers can optionally use classes that are defined in the **LMF extensions** (see relevant annexes). Developers shall define a data category selection (DCS) as specified for **LMF data category selection procedures** (5.4).

5.2 LMF core package

The LMF core package is a metamodel that provides a flexible basis for building LMF models and extensions, see Figure 1.

5.2.1 Lexical Resource class

Lexical Resource is a class representing the entire resource. *Lexical Resource* occurs once and only once. The *Lexical Resource* instance is a container for one or more lexicons.

5.2.2 Global Information class

Global Information is a class representing administrative information and other general attributes. There is an aggregation relationship between the *Lexical Resource* class and the *Global Information* class in that the latter describes the administrative information and general attributes of the entire resource. The *Global Information* class does not allow subclasses.

The *Global Information* instance must contain at least the following attribute:

— /language coding/ This attribute specifies which standard is used in order to code the language names within the whole *Lexical Resource* instance.

The *Global Information* instance may contain the following attributes:

— /script coding/ This attribute specifies which standard is used in order to code the script names within the whole *Lexical Resource* instance;

— /character coding/ This attribute specifies which Unicode version is used within the whole *Lexical Resource* instance.

NOTE Other standard related precisions may be specified on the *Global Information* instance.

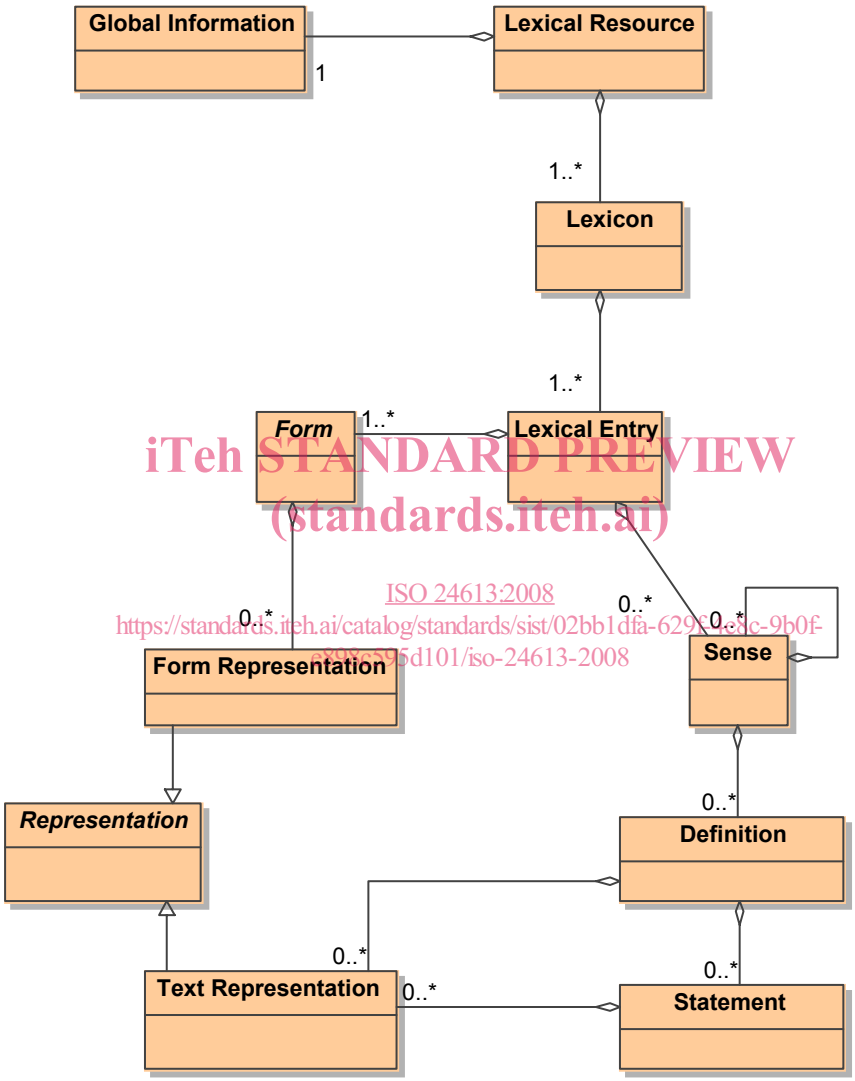

Figure 1 — LMF core package

5.2.3 Lexicon class

Lexicon is a class containing all the lexical entries of a given language within the entire resource. A *Lexicon* instance must contain at least one lexical entry. The *Lexicon* class does not allow subclasses.

5.2.4 Lexical Entry class

Lexical Entry is a class representing a lexeme in a given language. The *Lexical Entry* is a container for managing the *Form* and *Sense* classes. Therefore, the *Lexical Entry* manages the relationship between the forms and their related senses. A *Lexical Entry* instance can contain one to many different forms, and can have from zero to many different senses. The *Lexical Entry* class does not allow subclasses.

5.2.5 Form class

Form class is an abstract class representing a lexeme, a morphological variant of a lexeme or a morph. The *Form* class manages one or more orthographical variants of the abstract *Form* as well as data categories that describe the attributes of the word form (e.g. lemma, pronunciation, syllabification). The *Form* class allows subclasses.

5.2.6 Form Representation class

Form Representation is a class representing one variant orthography of a *Form*. When there is more than one variant orthography, the *Form Representation* class contains a Unicode string representing the *Form* as well as, if needed, the unique attribute-value pairs that describe the specific language, script, and orthography.

5.2.7 Representation class

Representation is an abstract class representing a Unicode string as well as, if needed, the unique attribute-value pairs that describe the specific language, script, and orthography. The *Representation* class allows subclasses.

5.2.8 Sense class

Sense is a class representing one meaning of a lexical entry. The *Sense* class allows subclasses. The *Sense* class also allows for hierarchical senses in that one sense may be more specific than another sense of the same lexical entry.

5.2.9 Definition class

Definition is a class representing a narrative description of a sense. It is displayed for human users to facilitate their understanding of the meaning of a *Lexical Entry* and is not meant to be processable by computer programs. A *Sense* instance can have zero to many definitions. Each *Definition* instance may be associated with zero to many *Text Representation* instances in order to manage the text definition in more than one language or script. The narrative description can be expressed in a different language and/or script than the one for the *Lexical Entry* instance.

EXAMPLE In a *Lexical Entry* for *abbess*, the narrative description may be *woman who is in charge of a convent*.

5.2.10 Statement class

Statement is a class representing a narrative description and refines or complements *Definition*. A *Definition* instance can have zero to many *Statement* instances.

NOTE A full example is given in WordNet context in Annex H.

5.2.11 Text Representation class

Text Representation is a class representing one textual content of *Definition* or *Statement*. When there is more than one variant orthography, the *Text Representation* class contains a Unicode string representing the textual content as well as the unique attribute-value pairs that describe the specific language, script, and orthography.