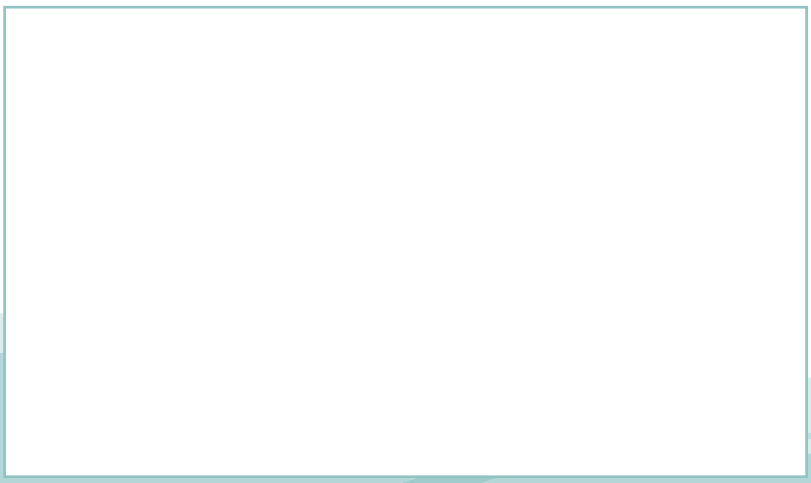
SVENSK STANDARD
SS-ISO 24613-1:2019

Fastställd/Approved: 2019-10-01
Utgåva/Edition: 1
Språk/Language: engelska/English
ICS: 01.020; 35.240.30

SWEDISH STANDARDS INSTITUTE

Hantering av språkliga resurser – Strukturerat lexikonformat (LMF) – Del 1: Grundmodell (ISO 24613-1:2019, IDT)

Language resource management – Lexical markup framework (LMF) – Part 1: Core model (ISO 24613-1:2019, IDT)



Standarder får världen att fungera

SIS (Swedish Standards Institute) är en fristående ideell förening med medlemmar från både privat och offentlig sektor. Vi är en del av det europeiska och globala nätverk som utarbetar internationella standarder. Standarder är dokumenterad kunskap utvecklad av framstående aktörer inom industri, näringsliv och samhälle och befrämjar handel över gränser, bidrar till att processer och produkter blir säkrare samt effektiviserar din verksamhet.

Delta och påverka
Som medlem i SIS har du möjlighet att påverka framtida standarder inom ditt område på nationell, europeisk och global nivå. Du får samtidigt tillgång till tidig information om utvecklingen inom din bransch.

Ta del av det färdiga arbetet
Vi erbjuder våra kunder allt som rör standarder och deras tillämpning. Hos oss kan du köpa alla publikationer du behöver – allt från enskilda standarder, tekniska rapporter och standard-paket till handböcker och onlinetjänster. Genom vår webbtjänst e-nav får du tillgång till ett lättnavigerat bibliotek där alla standarder som är aktuella för ditt företag finns tillgängliga. Standarder och handböcker är källor till kunskap. Vi säljer dem.

Utveckla din kompetens och lyckas bättre i ditt arbete
Hos SIS kan du gå öppna eller företagsinterna utbildningar kring innehåll och tillämpning av standarder. Genom vår närhet till den internationella utvecklingen och ISO får du rätt kunskap i rätt tid, direkt från källan. Med vår kunskap om standarders möjligheter hjälper vi våra kunder att skapa verklig nytta och lönsamhet i sina verksamheter.

Vill du veta mer om SIS eller hur standarder kan effektivisera din verksamhet är du välkommen in på www.sis.se eller ta kontakt med oss på tel 08-555 523 00.

Standards make the world go round

SIS (Swedish Standards Institute) is an independent non-profit organisation with members from both the private and public sectors. We are part of the European and global network that draws up international standards. Standards consist of documented knowledge developed by prominent actors within the industry, business world and society. They promote cross-border trade, they help to make processes and products safer and they streamline your organisation.

Take part and have influence
As a member of SIS you will have the possibility to participate in standardization activities on national, European and global level. The membership in SIS will give you the opportunity to influence future standards and gain access to early stage information about developments within your field.

Get to know the finished work
We offer our customers everything in connection with standards and their application. You can purchase all the publications you need from us - everything from individual standards, technical reports and standard packages through to manuals and online services. Our web service e-nav gives you access to an easy-to-navigate library where all standards that are relevant to your company are available. Standards and manuals are sources of knowledge. We sell them.

Increase understanding and improve perception
With SIS you can undergo either shared or in-house training in the content and application of standards. Thanks to our proximity to international development and ISO you receive the right knowledge at the right time, direct from the source. With our knowledge about the potential of standards, we assist our customers in creating tangible benefit and profitability in their organisations.

If you want to know more about SIS, or how standards can streamline your organisation, please visit www.sis.se or contact us on phone +46 (0)8-555 523 00

Den internationella standarden ISO 24613-1:2019 gäller som svensk standard. Detta dokument innehåller den officiella engelska versionen av ISO 24613-1:2019.

The International Standard ISO 24613-1:2019 has the status of a Swedish Standard. This document contains the official English version of ISO 24613-1:2019.

Upplysningar om sakinnehållet i standarden lämnas av SIS, Swedish Standards Institute, telefon 08-555 520 00. Standarder kan beställas hos SIS som även lämnar allmänna upplysningar om svensk och utländsk standard.

Information about the content of the standard is available from the Swedish Standards Institute (SIS), telephone +46 8 555 520 00. Standards may be ordered from SIS, who can also provide general information about Swedish and foreign standards.

Denna standard är framtagen av kommittén för Språk och terminologi, SIS/TK 115.

Har du synpunkter på innehållet i den här standarden, vill du delta i ett kommande revideringsarbete eller vara med och ta fram andra standarder inom området? Gå in på www.sis.se - där hittar du mer information.

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 of 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 www.iso.org/iso/foreword.html.

The document was prepared by Technical Committee ISO/TC 37, *Language and terminology,* Subcommittee 4, *Language resource management*.

This first edition of ISO 24613-1, together with ISO 24613-2 to ISO 24613-6, cancels and replaces ISO 24613:2008, which has been technically revised.

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

The content has been entirely revised and subdivided into parts. Part 1, Core model, contains the body of the previous edition. New classes include LexiconInformation and GrammaticalInformation. The Representation class has been renamed the OrthographicRepresentation class. In addition, the OrthographicRepresentation subclasses, FormRepresentation and TextRepresentation, no longer are part of the core model, providing it with greater modeling flexibility. The LexicalEntry subclass now allows subclasses, providing improved extensibility and flexibility for modeling future parts. The addition of the CrossREF class and associated metadata provides a formal model for cross-reference design and implementation, closing a functional gap in the previous edition. A thoroughly revised description of data category allocation mechanisms and their relationship to generalization by typing provides a more incisive description of how these interdependent mechanisms enable flexible and extensible designs.

A list of all parts in the ISO 24613 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user’s national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

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.

LMF supports existing lexical resource models such as Genelex[3], the EAGLES International Standard for Language Engineering (ISLE)[4], Multilingual ISLE Lexical Entry (MILE) models[10], Text Encoding Initiative (TEI) guidelines[8], Ontolex[7], and the Language Base Exchange (LBX) serialization together with the U.S. Government Wordscape On-Line Dictionary system[5].

LMF uses UML modeling processes[9]. The LMF core package describes the basic hierarchy of information of a lexical entry, including information on the word 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 in additional LMF parts (see Annex A for data category examples);

— 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.

LMF parts 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.

The parts currently in or planned for the new organization of ISO 24613 include *Part 1: Core model,* Part 2: *Machine readable dictionary (MRD) model*, Part 3: *Diachrony-etymology*, Part 4: *TEI serialization*, *Part 5: LBX serialization*, and Part 6: *Syntax and semantics*.

The ISO 24613 series is designed to coordinate closely with ISO 16642[2].

Language resource management — Lexical markup framework (LMF) —

Part 1:
Core model

1 Scope

This document describes the core model of the lexical markup framework (LMF)l, a metamodel for representing data in monolingual and multilingual lexical databases used with computer applications.

LMF provides mechanisms that allow the development and integration of a variety of electronic lexical resource types.

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

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

3 Terms and definitions

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

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

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

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

3.1
data category
DC
<lexical markup framework> elementary descriptor used in a linguistic description or annotation scheme

3.2
word form
instance of a word, multi-word expression, root, stem, or morpheme

3.3
grammatical feature
property associated with a *word form* (3.2) to describe one of its grammatical attributes

EXAMPLE /grammatical gender/

3.4
lemma
lemmatized form
canonical form
conventional *word form* (3.2) chosen to represent a *lexeme* (3.5)

Note 1 to entry: In many European languages, the lemma is usually the /singular/ for a noun 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 sometimes considered as being the third person singular with the accomplished aspect, in other approaches it is considered as being the root.

3.5
lexeme
abstract unit generally associated with a set of *word forms* (3.2) sharing a common meaning

[SOURCE: ISO 24613:2008, 3.25, modified – "forms" replaced with "word forms".]

3.6
lexical resource
lexical database
database consisting of one or several *lexicons* (3.7)

3.7
lexicon
resource comprising lexical entries for one or several languages

Note 1 to entry: A special language lexicon or a lexicon prepared for a specific NLP application can comprise a specific subset of a language.

3.8
multiword expression
MWE
lexeme (3.5) made up of a sequence of two or more lexemes that has properties that may not be predictable from the properties of the individual lexemes or their normal mode of combination

EXAMPLE "To kick the bucket", an idiomatic expression which means to die rather than to hit a bucket with one's foot. An idiomatic expression is a subtype of MWE whose properties are not predictable from the properties of the individual lexemes.

Note 1 to entry: 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 (3.13).

3.9
natural language processing
NLP
field covering knowledge and techniques involved in the processing of linguistic data by a computer

3.10
orthography
way of spelling or writing *lexemes* (3.5) that conforms to a conventionalized use

Note 1 to entry: Usually, the notion of orthography covers standardized spellings of alphabetic languages, such as standard UK or US English, or reformed German spelling, as well as hieroglyphic or syllabic writing systems. For the purpose of this standard, we also subsume variations such as transliterations of languages in non-native scripts, stenographic renderings, or representations in the International Phonetic Alphabet under the notion of orthography.

3.11
part of speech
lexical category
word class
category assigned to a *lexeme* (3.5) based on its grammatical properties

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

3.12
script
set of graphic characters used for the written form of one or more languages

EXAMPLE Hiragana, Katakana, Latin and Cyrillic.

Note 1 to entry: 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.

[SOURCE: ISO/IEC 10646:2017 3.50, modified – Example and Note 1 to entry added]

4 Key standards used by LMF

4.1 Unicode

LMF is Unicode-compliant and presumes that all data are used according to the Unicode character encodings.

4.2 Language coding

Language identifiers used in LMF-compliant resources shall conform to criteria specified in the ISO 639 series 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. The current edition of IETF Best Common Practices (BCP) 47[6] should be consulted.

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 Unified modeling language (UML)

LMF complies with the specifications and modeling principles of UML as defined by the Object Management Group (OMG)[9], 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 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 describe the semantics of the modeling elements and 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** (5.3), and classes that are defined in other **LMF parts** or classes derived from any of these referenced

classes using documented LMF processes for class inheritance. Developers shall define a data category selection (DCS) as specified for **LMF data category selection procedures** (5.2.3 and 5.2.4).

5.2 Class inheritance and data category selection procedures

5.2.1 Class inheritance

LMF specifies constraints on which classes allow subclasses.

5.2.2 LMF attributes

UML models such as LMF are populated or further described by UML attributes, which provide information about specific properties or characteristics associated with the model. All LMF attributes are complex data categories. For a given class, all attributes are different. Each value of an attribute is either a simple data category or a Unicode string. Each attribute has only one value.

5.2.3 Data category selection (DCS)

In the broadest sense, a data category selection can comprise all the data categories used by a given domain in the field of language resources. A DCS can also list and describe the set of data categories that can be used in a given LMF lexicon. The DCS also describes constraints on how the data categories are mapped to specific classes.

5.2.4 User-defined data categories

Lexicon creators can define a set of new data categories to cover data category concepts that are needed and that are not available.

5.3 LMF core package

5.3.1 General

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

NOTE Each word in a class name begins with a capital letter with no intervening spaces or punctuation. This practice is not required by UML, but generally conforms with most UML documentation.

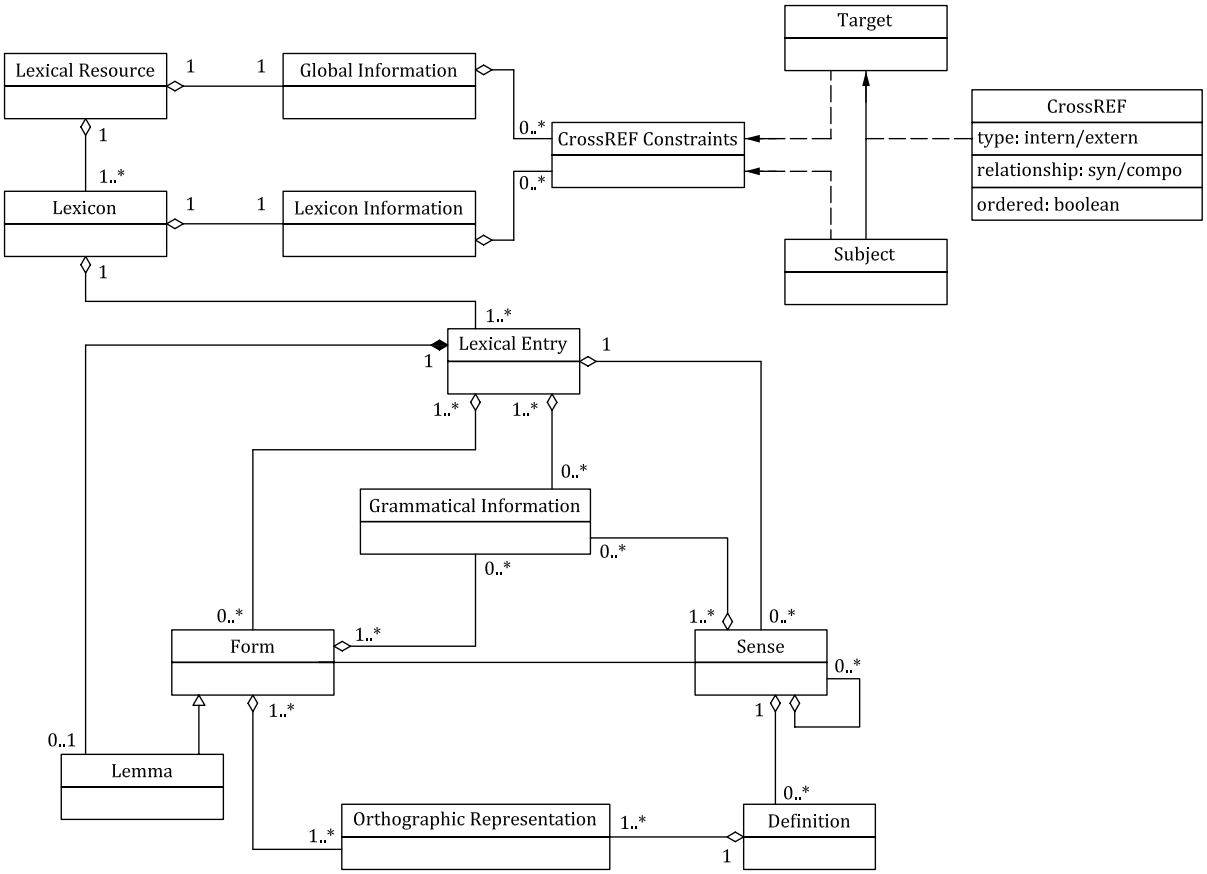


Figure 1 — LMF core package

5.3.2 LexicalResource class

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

5.3.3 GlobalInformation class

GlobalInformation is a class representing administrative information and other general attributes. There is a one-to-one aggregate association between the Lexicon class and the GlobalInformation class in that the latter describes the administrative information and general attributes of the entire resource. The GlobalInformation class does not allow subclasses.

The GlobalInformation instance shall contain at least the following attributes:

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

The GlobalInformation instance can contain the following attributes:

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

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

NOTE Other standard related precisions can be specified on the GlobalInformation instance.