INTERNATIONAL STANDARD

ISO 24613-2

First edition
2020-07

Language resource management — Lexical markup framework (LMF) —

Part 2: Machine-readable dictionary (MRD) model

Gestion des ressources linguistiques — Cadre de balisage lexical (LMF) —

Partie 2: Modèle de dictionnaire lisible par ordinateur (MRD)

Reference number
ISO 24613-2:2020(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword ... **iv**
Introduction ... **v**
1 Scope ... **1**
2 Normative references ... **1**
3 Terms and definitions ... **1**
4 Key standards used by LMF ... **1**
5 The machine-readable dictionary (MRD) model ... **1**
5.1 General ... 1
5.2 MRD class model ... 2
5.2.1 Set of classes ... 2
5.2.2 Class selection and multiplicity ... 2
5.2.3 Generalization ... 3
5.2.4 Object realization ... 3
5.3 Data category selection and class population ... 3
5.4 CrossREF allocation ... 3
5.5 Form subclasses ... 4
5.5.1 WordForm class ... 4
5.5.2 Lemma class ... 4
5.5.3 Stem class ... 4
5.5.4 WordPart class ... 4
5.5.5 RelatedForm class ... 4
5.6 FormRepresentation class ... 4
5.7 TextRepresentation class ... 5
5.8 Translation class ... 5
5.9 Example class ... 5
5.10 SubjectField class ... 5
5.11 Bibliography class ... 5
5.12 Multiword Expression (MWE) Analysis ... 6
Annex A (informative) **Data category examples** ... **7**
Annex B (informative) **Machine-readable dictionary examples** ... **9**
Bibliography ... **21**

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.

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

This first edition of ISO 24613-2, together with ISO 24613-1:2019, ISO 24613-3[1)], ISO 24613-4[1)], ISO 24613-5[1)], ISO 24613-6[2)] and ISO 24613-7[2)], cancels and replaces ISO 24613:2008, which has been divided into several parts and technically revised.

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

This edition merges two normative annexes from the previous edition, Annex A, Morphology extension, and Annex C, Machine-readable dictionary extension, providing a more cohesive description of the key structures (classes and associations) found in that edition. The cross-reference (CrossREF) model introduced in Part 1, Core model, of this edition, provides a new capability for correlating lexical features across different form and sense classes. In addition, the CrossREF model has replaced the ListOfComponents and Component classes, enabling a more extensible and flexible capability for managing multiword expressions. The metamodel of generalization by typing introduced in Part 1 provides a more rigorous and unambiguous framework for applying LMF modelling mechanisms in ways that enable greater editorial freedom and support the comparison of different LMF conformant designs. This edition has kept most of the informative examples found in the previous edition (deleting only a few redundant examples) and has added new examples to illustrate new modelling features. There have been some class name changes (e.g. OrthographicRepresentation for Representation and Translation for Equivalent), but no changes in the underlying concepts of the previously existing classes.

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.

1) Under preparation.

2) Planned.

Introduction

The ISO 24613 series is based upon the definition of an implementation-independent metamodel combining a core model and additional models that onomasiological (form-oriented) and semasiological (concept-oriented) lexical content can take.

It provides guidelines for various implementation use cases, and where appropriate describes LMF compliant serializations that fit various application contexts.

This document extends ISO 24613-1, the LMF core model, through the use of the processes and mechanisms described in ISO 24613-1. The objective is to enable flexible design methods to support the development of machine-readable dictionaries for different purposes while enabling cross-comparisons of different designs and a basis for developing assessments of standards conformance. The scope of supported design goals ranges from simple to complex human-oriented MRDs, both monolingual and bilingual, lexicons that support conceptual-lexical systems through links with ontological resources, rigorously constrained lexicons for supporting machine processes, and lexicons that provide an extensional description of the morphology of lexical entries. Since this document is based on ISO 24613-1, the LMF core model, it is designed to interchange data with other parts of the ISO 24613 series where applicable.

Language resource management — Lexical markup framework (LMF) —

Part 2: Machine-readable dictionary (MRD) model

IMPORTANT — The electronic file of this document contains colours which are considered to be useful for the correct understanding of the document. Users should therefore consider printing this document using a colour printer.

1 Scope

This document describes the machine-readable dictionary (MRD) model, a metamodel for representing data stored in a variety of electronic dictionary subtypes, ranging from direct support for human translators to support for machine processing.

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 24613-1, *Language resource management — Lexical markup framework (LMF) — Part 1: Core model*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 24613-1 apply.

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

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

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

4 Key standards used by LMF

The key standards applicable to this document are described in ISO 24613-1, the LMF core model.

5 The machine-readable dictionary (MRD) model

5.1 General

The MRD model is represented by UML classes, associations among the classes (the structure), sets of data categories (attribute-value pairs), and links (cross-references). Subclauses 5.2 through 5.12 describe each of these features, their interdependencies, and their implementation.

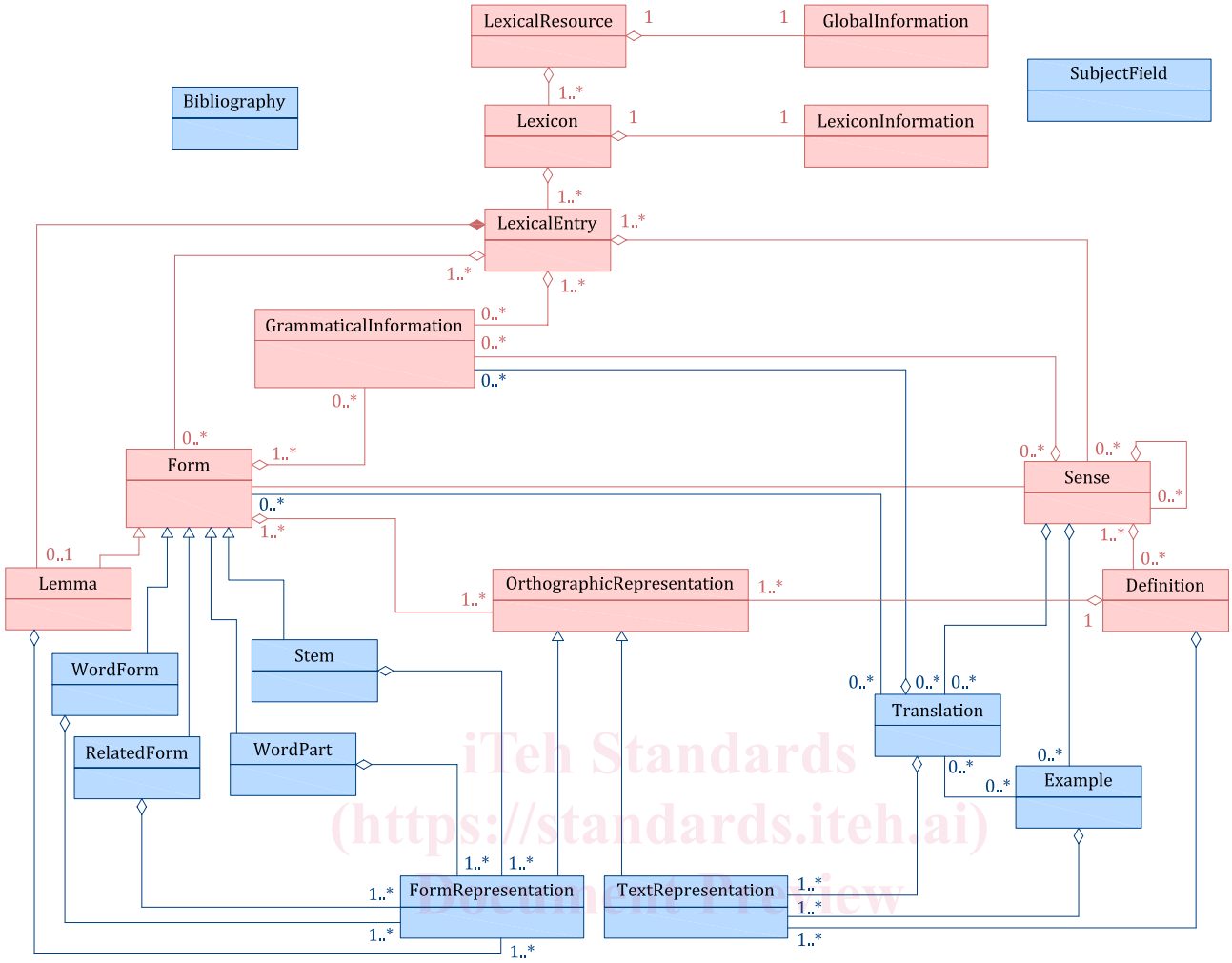

Figure 1 — MRD class model

5.2 MRD class model

5.2.1 Set of classes

The classes defined in ISO 24613-1, the LMF core model, that are used in the MRD extension include LexicalResource, GlobalInformation, Lexicon, LexiconInformation, GrammaticalInformation, LexicalEntry, Lemma, Form, Sense, Definition, OrthographicRepresentation, and principles for applying the CrossREF class. These classes, together with the associations and constraints described in ISO 24613-1, are applicable to the design of MRD. New classes introduced in this document include WordForm, Stem, WordPart, RelatedForm, Translation, Example, FormRepresentation, TextRepresentation, Bibliography and SubjectField.

5.2.2 Class selection and multiplicity

The sets of classes shown in the model in Figure 1 can support a wide range of design objectives. A specific design objective can require all or only some of the classes shown in the above model and can require as well the creation of new subclasses. The recommended first step in the creation of a model for a specific design objective (e.g. a bilingual dictionary) should be the selection and possible exclusion of classes contained in the class model and the application of desired multiplicities to the class associations as required by the model and the design goals (the optional classes in the model have a minimum cardinality of zero). The developer can create new subclasses, as needed, using the mechanisms described in ISO 24613-1, the LMF core model. The selected classes and their associations

provide the structure and nodes (classes) appropriate for the intended lexical design. The classes and subclasses are described in detail below (see 5.5 to 5.11).

EXAMPLE

— Certain classes of MRD, such as monolingual and bilingual dictionaries, generally require a Sense class instantiation.

— Certain classes of MRD, such as concept hierarchies, do not necessarily require a Form class instantiation.

— Certain classes of MRD, such as orthographic dictionaries and extensional morphologies do not necessarily require a Sense class instantiation.

— Certain classes of MRD, such as extensional morphologies, can provide constraints on the attributes managed by the RelatedForm class.

NOTE The purpose of the MRD morphology extension is to provide the mechanisms to support the development of lexicons that have an extensional description of the morphology of lexical entries in which all relevant inflections or derivations of a lemma are included.

5.2.3 Generalization

Figure 1 illustrates the use of generalization (typing) through the Form class (superclass) and its subclasses, Lemma, WordForm, Stem, and WordPart, and OrthographicRepresentation (superclass) and its subclasses, FormRepresentation and TextRepresentation. The typing mechanism describes how to allocate specific sets of data categories, associations, multiplicities, and cross-references to subclasses (e.g. Lemma) in order to redefine the superclass. ISO 24613-1 provides a more complete description of typing.

NOTE The subclasses shown in Figure 1 are available for use in LMF compliant designs, but are not exhaustive, since LMF allows the creation of additional subclasses. The lexicon designer specifies what sets of features are available in form features.

5.2.4 Object realization

LMF provides examples of object models (see Annex B), but does not provide an in-depth description of the overall methodologies for developing the object models, since those processes are heavily dependent on the choice of model serialization (e.g. XML, JSON). Different serializations can require different design approaches and impose limitations on how the object can be modelled.

EXAMPLE XML provides a number of structural models for implementing XML schemas. Within the framework of these models, a lexicon designer could implement UML classes as XML elements or a combination of an XML element and attributes. For example, a designer could instantiate the Lemma class as a <Lemma> element or a <form type="lemma"> element-attribute combination. These object modelling choices use selective class and data category allocations to implement object designs that are strongly dependent on the structures and methods of the chosen serialization.

5.3 Data category selection and class population

Data category selection can include all or a subset of data categories used by a given domain. Examples of data categories and their allocations are listed in Annex A. Where needed, the lexicon developer can create new data categories that are not listed in the annex.

5.4 CrossREF allocation

Figure 1 shows links (cross-references) between the Form and Sense and the Form and Translation classes. The principles for modelling cross-references are described in ISO 24613-1, the LMF core model. The CrossREF class is specifically allowed for the LexicalEntry class, the Lemma class, the WordForm class, the WordPart class, the Sense class, and the Sense class children. The lexicon designer should consider using cross-references with the RelatedForm class. The use of data categories to provide

information about the CrossREF features (e.g. internal reference, external reference, type of ID, lexical type, syntactic type, or semantic type) is a best practice.

EXAMPLE A WordPart that contains the suffix component of a Lemma can be cross-referenced with the LexicalEntry that contains that suffix as the Lemma, or a Sense can be cross-referenced with a broader Sense contained in a different LexicalEntry, or an authentic Quote can be cross-referenced with a document that contains the Quote.

NOTE The range of data categories describing CrossREF features is potentially quite broad and could be used to support references to audio, video, and other types of metadata relevant for lexical resources.

5.5 Form subclasses

5.5.1 WordForm class

WordForm is a Form subclass containing a word form, such as an inflected form, that a lexeme can take when used in a sentence or a phrase. The WordForm class is in a zero-to-many aggregate association with the LexicalEntry class (inheriting the Form multiplicity). The WordForm class can manage simple lexemes, compounds, multi-word expressions, and sub-lexemes such as affixes and roots.

5.5.2 Lemma class

Lemma is a Form subclass representing a lexeme or sub-lexeme used to designate the LexicalEntry (part of the Form-Sense paradigm). The Lemma class is in a zero-to-one aggregate association with the LexicalEntry class that overrides the multiplicity inherited from the Form class (see ISO 24613-1 for a more complete description of the Lemma).

5.5.3 Stem class

Stem is a Form subclass containing a stem or root. The Stem class can be typed as a specific type of stem or root (e.g. type="arabicRoot"). The Stem class is in a zero-to-one aggregate association with the LexicalEntry class (overriding the multiplicity inherited from the Form class).

5.5.4 WordPart class

WordPart is a Form subclass representing sub-lexeme parts other than the stem or root (e.g. affix, prefix, suffix). The WordPart class is in a zero-to-many aggregate association with the LexicalEntry class.

5.5.5 RelatedForm class

RelatedForm is a Form subclass containing a word form or a morph that is typical of run-on entries in print dictionaries. The RelatedForm has a different Sense than the Lemma and can be considered a candidate for eventual inclusion in a different LexicalEntry object when realized in a lexical database. The RelatedForm can be related to the Lemma in a variety of ways (e.g. synonym, cross-reference, multi-word expression, idiom). The RelatedForm class is in a zero-to-many aggregate association with the LexicalEntry class and can contain a recursive cross-reference to the LexicalEntry class, which would be realized as a link to a different LexicalEntry object when instantiated in a lexical database. The RelatedForm class can be typed (generalization) using data categories.

EXAMPLE A developer possibly wants to use the RelatedForm class for a multi-word expression (e.g. *United States*) that contains a component form of a Lemma (e.g. *united*). The design goal could be to preserve the format of the original source material, or to provide immediate user support while developing an improved lexicon that includes /united/ and /United States/ as separate entries.

5.6 FormRepresentation class

FormRepresentation is an OrthographicRepresentation subclass that contains the text literals and metadata (e.g. pronunciation, hyphenation, xml:lang, script) for a Lemma, WordForm, or other subclass

