INTERNATIONAL STANDARD

ISO 24613-1

First edition
2019-06

Language resource management — Lexical markup framework (LMF) —

Part 1: Core model

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

Partie 1: Modèle de base

This document is a preview generated by EVS

Reference number
ISO 24613-1:2019(E)

© ISO 2019

This document is a preview generated by EVS

COPYRIGHT PROTECTED DOCUMENT

© ISO 2019

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

 © ISO 2019 – All rights reserved

Contents

Page

Foreword ... **iv**
Introduction ... **v**
1 Scope ... **1**
2 Normative references ... **1**
3 Terms and definitions ... **1**
4 Key standards used by LMF ... **3**
4.1 Unicode ... 3
4.2 Language coding ... 3
4.3 Script coding ... 3
4.4 Unified modeling language (UML) ... 3
5 The LMF model ... **3**
5.1 Introduction ... 3
5.2 Class inheritance and data category selection procedures ... 4
5.2.1 Class inheritance ... 4
5.2.2 LMF attributes ... 4
5.2.3 Data category selection (DCS) ... 4
5.2.4 User-defined data categories ... 4
5.3 LMF core package ... 4
5.3.1 General ... 4
5.3.2 LexicalResource class ... 5
5.3.3 GlobalInformation class ... 5
5.3.4 Lexicon class ... 6
5.3.5 LexiconInformation class ... 6
5.3.6 LexicalEntry class ... 6
5.3.7 Form class ... 6
5.3.8 OrthographicRepresentation class ... 6
5.3.9 GrammaticalInformation Class ... 6
5.3.10 Sense class ... 7
5.3.11 Definition class ... 7
5.4 Cross reference (CrossREF) model ... 7
5.4.1 General ... 7
5.4.2 CrossREF and CrossREFConstraint classes ... 7
5.4.3 CrossREFConstraint class ... 7
5.5 Methods for data category selection and subclass creation ... 7
5.5.1 General ... 7
5.5.2 Generalization (typing) ... 8
5.5.3 Object instantiation ... 8
5.5.4 Design choices ... 8
5.5.5 Data categories for orthographic representation ... 9
5.5.6 Principles for model simplification ... 9
5.6 LMF extension use ... 9
5.6.1 General ... 9
5.6.2 Lexicon comparison ... 10
Annex A (informative) **Data category examples** ... **11**
Bibliography ... **13**

This document is a preview generated by EVS

© ISO 2019 – All rights reserved

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.

This document is a preview generated by EVS

 © ISO 2019 – All rights reserved

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].

This document is a preview generated by EVS

© ISO 2019 – All rights reserved

This document is a preview generated by EVS

Language resource management — Lexical markup framework (LMF) —

Part 1: Core model

This document is a preview generated by EVS

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/

© ISO 2019 – All rights reserved