© ISO 2023 – All rights reserved

ISO/DIS PRF 24613-1:2023(E)

Date: 2023-03-06

ISO/TC 37/SC 4/WG 4

Secretariat: KATS

Date: 2023-11-13

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

iTeh Standards
(https://standards.iteh.ai)
Document Preview

ISO/PRF 24613-1
https://standards.iteh.ai/catalog/standards/sist/5aa02d04-5dab-4356-816c-06107b0ef4c8/iso-prf-24613-1

PROOF

© ISO 2023

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~~ copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: + 41 22 749 01 11

~~Email:~~ ~~copyright@iso.org~~
E-mail: copyright@iso.org
Website: www.iso.org

Published in Switzerland~~.~~

iTeh Standards
(https://standards.iteh.ai)
Document Preview

ISO/PRF 24613-1
https://standards.iteh.ai/catalog/standards/sist/5aa02d04-5dab-4356-816c-06107b0ef4c8/iso-prf-24613-1

Contents

Foreword vi
Introduction viii
Part 1: Core model 1
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 4
4.4 Unified modelling language 4
5 The LMF model 4
5.1 General 4
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 5
5.3.1 General 5
5.3.2 LexicalResource class 6
5.3.3 GlobalInformation class 6
5.3.4 Lexicon class 7
5.3.5 LexiconInformation class 7
5.3.6 LexicalEntry class 7
5.3.7 Form class 7
5.3.8 OrthographicRepresentation class 7
5.3.9 GrammaticalInformation class 7
5.3.10 Sense class 8
5.3.11 Definition class 8
5.4 Cross reference (CrossREF) model 8
5.4.1 General 8
5.4.2 CrossREF classes 8
5.4.3 CrossREFConstraint class 8
5.5 Methods for data category selection and subclass creation 8
5.5.1 General 8

iTeh Standards (https://standards.iteh.ai) Document Preview
ISO/PRF 24613-1
https://standards.iteh.ai/catalog/standards/sist/5aa02d04-5dab-4356-816c-06107b0ef4c8/iso-prf-24613-1

~~© ISO 2023 – All rights reserved~~

© ISO 2023 – All rights reserved

5.5.2 Generalization 9
5.5.3 Object instantiation 9
5.5.4 Design choices 10
5.5.5 Data categories for orthographic representation 10
5.5.6 Principles for model simplification 10
5.6 LMF extension use 11
5.6.1 General 11
5.6.2 Lexicon comparison 12
Annex A (informative) Data category examples 13
Bibliography 16

~~Foreword iv~~
~~Introduction vi~~
~~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 modelling language 3~~
~~5 The LMF model 3~~
~~5.1 General 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 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~~

iTeh Standards
(https://standards.iteh.ai)
Document Preview
ISO/PRF 24613-1
https://standards.iteh.ai/catalog/standards/sist/5aa02d04-5dab-4356-816c-06107b0ef4c8/iso-prf-24613-1

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) 7
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 12
Bibliography 15

iTeh Standards
(https://standards.iteh.ai)
Document Preview
ISO/PRF 24613-1
https://standards.iteh.ai/catalog/standards/sist/5aa02d04-5dab-4356-816c-06107b0ef4c8/iso-prf-24613-1

© ISO 2023 – All rights reserved

© ISO 2023 – 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~~document 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~~ISO draws attention to the possibility that ~~some of~~ the ~~elements~~implementation of this document may ~~be~~involve the ~~subject~~use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights~~.~~ in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. 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).~~

iTeh Standards
(https://standards.iteh.ai)
Document Preview

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~~ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

ISO 24613-1:2023
https://standards.iteh.ai/catalog/standards/sist/5aa02d04-5dab-4356-816c-06107b0ef4c8/iso-prf-24613-1

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

This second edition cancels and replaces the first edition (ISO 24613-1:2019), which has been technically revised.

The main changes are as follows:

— ~~—~~several changes have been made to ~~Figure 1~~Figure 1 "LMF core package", as follows:

  — ~~—~~the OrthographicRepresentation class associations with the Form and Definition classes previously had a cardinality of 1 to 1, which did not correctly represent the intent of the UML model; the revision of the cardinality to 1 to 0..* in each case now provides a correct model;

  — ~~—~~the type: intern/extern attribute-value pair is no longer included in the CrossREF class since it described linking processes relevant for implementations, not associations relevant for a metamodel;

  — ~~—~~the full names relationship values in the CrossREF class, "synonym/composition" replace the abbreviations, "syn/compo";

— —the class names in Figure 1Figure 1 are now harmonized with the LMF style;

— —relevant information has been moved from the tables in ISO 24613-2:2020 to Table A.1,Table A.1, meaning that the latter now contains more complete examples of values and attributes allocated to classes first introduced in this document.

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.

iTeh Standards
(https://standards.iteh.ai)
Document Preview
ISO/PRF 24613-1
https://standards.iteh.ai/catalog/standards/sist/5aa02d04-5dab-4356-816c-06107b0ef4c8/iso-prf-24613-1

Introduction

Optimizing the production, maintenance and extension of electronic lexical resources is one of the crucial aspects impacting human language technologies (HLTs) 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 instances, 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,[5] the EAGLES International Standard for Language Engineering (ISLE),[6] Multilingual ISLE Lexical Entry (MILE) models,[12] Text Encoding Initiative (TEI) guidelines,[10] Ontolex[9] and the Language Base Exchange (LBX) serialization together with the US Government Wordscape On-Line Dictionary system.[7]

LMF uses unified modelling language (UML) modelling processes.[11] 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. 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 that describe 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 ISO 24613 series is designed to coordinate closely with ISO 16642[3].

iTeh Standards (https://standards.iteh.ai) Document Preview
ISO/PRF 24613-1
https://standards.iteh.ai/catalog/standards/sist/5aa02d04-5dab-4356-816c-06107b0ef4c8/iso-prf-24613-1

© ISO 2023 – All rights reserved
viii © ISO 2023 – All rights reserved

Language resource management — Lexical markup framework (LMF) —

Part 1: Core model

1 Scope

This document establishes the core model of the lexical markup framework (LMF), a metamodel for representing data in monolingual and multilingual lexical resources 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, *Code for individual languages and language groups*

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

iTeh Standards (https://standards.iteh.ai) Document Preview

3 Terms and definitions

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

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

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

— IEC Electropedia: available at ~~https://www.electropedia.org/~~https://www.electropedia.org/

3.1
data category
DC
class of data items that are closely related from a formal or semantic point of view

EXAMPLE: /part of speech/, /subject field/, /definition/.

Note 1 to entry: A data category can be viewed as a generalization of the notion of a field in a database.

Note 2 to entry: In running text, such as this document, data category names are enclosed in forward slashes (e.g. /part of speech/).

© ISO 2023 – All rights reserved

[SOURCE: ISO 30042:2019, 3.8, modified – admitted term "DC" added]

3.2
word form
instantiation of a lexeme in a syntactic context

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

EXAMPLE grammaticalGender.

3.4
lemma
lemmatized form
canonical form
a *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. In other approaches it is considered as being the root.

iTeh Standards
(https://standards.iteh.ai)
Document Preview

3.5
lexeme
abstract unit generally associated with a set of *word forms* (3.2) sharing common properties, such as morphologic, morpho-syntactic, semantic, or phonetic properties.

ISO/PRF 24613-1
https://standards.iteh.ai/catalog/standards/sist/5aa02d04-5dab-4356-816c-06107b0ef4c8/iso-prf-24613-1

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 natural language processing 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 are not necessarily 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.

© ISO 2023 – All rights reserved
2 © ISO 2023 – All rights reserved