
**Upravljanje jezikovnih virov - Ogrodje za označevanje leksikonov (LMF) - 5. del:
Serializacija leksikalne osnovne izmenjave (LBX)**

Language resource management -- Lexical markup framework (LMF) - Part 5: Lexical
base exchange (LBX) serialization

iTeh STANDARD PREVIEW

Gestion des ressources linguistiques -- Cadre de balisage lexical (LMF) - Partie 5:
Sérialisation de l'échange de bases lexicales (LBX)

[oSIST ISO/DIS 24613-5:2021](https://standards.iteh.ai/catalog/standards/sist/24613-5-2021/24613-5-2021)

Ta slovenski standard je istoveten z: ISO/DIS 24613-5

<https://standards.iteh.ai/catalog/standards/sist/24613-5-2021/24613-5-2021>
[a70064c55cd5/osist-iso-dis-24613-5-2021](https://standards.iteh.ai/catalog/standards/sist/24613-5-2021/24613-5-2021)

ICS:

01.020	Terminologija (načela in koordinacija)	Terminology (principles and coordination)
01.140.20	Informacijske vede	Information sciences
35.240.30	Uporabniške rešitve IT v informatiki, dokumentiranju in založništvu	IT applications in information, documentation and publishing

oSIST ISO/DIS 24613-5:2021**en,fr**

iTeh STANDARD PREVIEW **(standards.iteh.ai)**

[oSIST ISO/DIS 24613-5:2021](https://standards.iteh.ai/catalog/standards/sist/2ed2a4d9-6480-492c-bed9-a70064c55cd5/osist-iso-dis-24613-5-2021)

<https://standards.iteh.ai/catalog/standards/sist/2ed2a4d9-6480-492c-bed9-a70064c55cd5/osist-iso-dis-24613-5-2021>

DRAFT INTERNATIONAL STANDARD

ISO/DIS 24613-5

ISO/TC 37/SC 4

Secretariat: KATS

Voting begins on:
2020-12-21Voting terminates on:
2021-03-15

Language resource management — Lexical markup framework (LMF) —

Part 5: Lexical base exchange (LBX) serialization

ICS: 01.020

iTeh STANDARD PREVIEW
(standards.iteh.ai)

[oSIST ISO/DIS 24613-5:2021](https://standards.iteh.ai/catalog/standards/sist/2ed2a4d9-6480-492c-bed9-a70064c55cd5/osist-iso-dis-24613-5-2021)<https://standards.iteh.ai/catalog/standards/sist/2ed2a4d9-6480-492c-bed9-a70064c55cd5/osist-iso-dis-24613-5-2021>

THIS DOCUMENT IS A DRAFT CIRCULATED FOR COMMENT AND APPROVAL. IT IS THEREFORE SUBJECT TO CHANGE AND MAY NOT BE REFERRED TO AS AN INTERNATIONAL STANDARD UNTIL PUBLISHED AS SUCH.

IN ADDITION TO THEIR EVALUATION AS BEING ACCEPTABLE FOR INDUSTRIAL, TECHNOLOGICAL, COMMERCIAL AND USER PURPOSES, DRAFT INTERNATIONAL STANDARDS MAY ON OCCASION HAVE TO BE CONSIDERED IN THE LIGHT OF THEIR POTENTIAL TO BECOME STANDARDS TO WHICH REFERENCE MAY BE MADE IN NATIONAL REGULATIONS.

RECIPIENTS OF THIS DRAFT ARE INVITED TO SUBMIT, WITH THEIR COMMENTS, NOTIFICATION OF ANY RELEVANT PATENT RIGHTS OF WHICH THEY ARE AWARE AND TO PROVIDE SUPPORTING DOCUMENTATION.

This document is circulated as received from the committee secretariat.



Reference number
ISO/DIS 24613-5:2020(E)

© ISO 2020

iTeh STANDARD PREVIEW (standards.iteh.ai)

[oSIST ISO/DIS 24613-5:2021](https://standards.iteh.ai/catalog/standards/sist/2ed2a4d9-6480-492c-bed9-a70064c55cd5/osist-iso-dis-24613-5-2021)

<https://standards.iteh.ai/catalog/standards/sist/2ed2a4d9-6480-492c-bed9-a70064c55cd5/osist-iso-dis-24613-5-2021>



COPYRIGHT PROTECTED DOCUMENT

© ISO 2020

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
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword	v
1 Scope	1
2 Normative references	1
3 Terms and definitions	1
4 General requirements	1
5 Serialization of the LMF core model (ISO 24613-1)	2
5.1 Implementing the LexicalResource class	2
5.2 Implementing the GlobalInformation class	2
5.3 Implementing the Lexicon class	3
5.4 Implementing the LexiconInformation class	3
5.5 Implementing the LexicalEntry class	4
5.6 Implementing the OrthographicRepresentation class	5
5.7 Implementing the Form class	6
5.7.1 Form class	6
5.7.2 Lemma class	6
5.8 Implementing the GrammaticalInformation class	6
5.9 Implementing the Sense class	7
5.10 Implementing the Definition class	7
5.11 Implementing CrossREF class	8
6 Serialization of the MRD extension (ISO 24613-2)	9
6.1 Implementing OrthographicRepresentation for MRD	9
6.2 Implementing Form representations for the Form subclasses	9
6.3 Classes derived from the Form class	10
6.3.1 General principles	10
6.3.2 Implementing the WordForm class	10
6.3.3 Implementing the Stem class	11
6.3.4 Implementing the WordPart class	11
6.3.5 Implementing the RelatedForm class	12
6.3.6 Implementing the TextRepresentation class	13
6.3.7 Implementing the Translation class	14
6.3.8 Implementing the Example class	14
6.4 Implementing the SubjectField class	14
6.5 Implementing the Bibliography class	15
7 Implementing the CrossREF mechanism to refer to external media files	15
8 Implementing the classes from the etymological extension (ISO 24613-3)	15
8.1 Implementing the Etymology class	15
8.2 Implementing the Etymon class	15
8.2.1 Referencing forms in an etymon	16
8.2.2 Representing the meaning of an etymon	16
8.2.3 Representing the language of an etymon	16
8.2.4 Dating an etymon	16
8.2.5 Providing sources associated with an etymon	16
8.3 Implementing the EtyLink class	16
8.4 Implementing the CognateSet class	17
8.5 Implementing the Cognate class	17
9 Additional mechanisms	18
9.1 Overview	18
9.2 XML feature structure implementation	18
9.3 Representing various labels with <LBL>	18
9.4 Providing rendering information with the @rend attribute	18
Annex A (informative) LBX data category selection	19

ISO/DIS 24613-5:2020(E)

Annex B (informative) LBX feature structure implementation	23
Annex C (informative) LBX examples for applying LBX serialization	26
Bibliography	31

iTeh STANDARD PREVIEW (standards.iteh.ai)

[oSIST ISO/DIS 24613-5:2021](https://standards.iteh.ai/catalog/standards/sist/2ed2a4d9-6480-492c-bed9-a70064c55cd5/osist-iso-dis-24613-5-2021)

<https://standards.iteh.ai/catalog/standards/sist/2ed2a4d9-6480-492c-bed9-a70064c55cd5/osist-iso-dis-24613-5-2021>

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 24613-5, together with ISO 24613-1 to -4, cancels and replaces ISO 24613:2008, which has been technically revised.

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

— entire revision of the content and its subdivisions.

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 STANDARD PREVIEW **(standards.iteh.ai)**

[oSIST ISO/DIS 24613-5:2021](https://standards.iteh.ai/catalog/standards/sist/2ed2a4d9-6480-492c-bed9-a70064c55cd5/osist-iso-dis-24613-5-2021)

<https://standards.iteh.ai/catalog/standards/sist/2ed2a4d9-6480-492c-bed9-a70064c55cd5/osist-iso-dis-24613-5-2021>

Language resource management — Lexical markup framework (LMF) —

Part 5: Lexical base exchange (LBX) serialization

1 Scope

This document describes the serialization of the LMF model defined as an XML model derived from the LBX schema and compliant with the W3C XML schema. This serialization covers the classes, data categories, and mechanisms of ISO 24613-1 (Core model), ISO 24613-2 (Machine-readable dictionary (MRD) model), and ISO 24613-3 (Etymological extension).

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.

BCP 47 Tags for Identifying Languages. A. Phillips; M. Davis. IETF. September 2009. IETF Best Current Practice. URL: <https://tools.ietf.org/html/bcp47>

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

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

ISO 24613-2, *Language resource management — Lexical markup framework (LMF) — Part 2: Machine-readable dictionary (MRD) model*

ISO 24613-3, *Language resource management — Lexical markup framework (LMF) — Part 3: Etymological extension*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 24613-1 and in ISO 24613-3 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 General requirements

This document aims at providing constructs for each LMF class from ISO 24613-1 (Core model), ISO 24613-2 (MRD extension), and ISO 24613-3 (Etymological extension). It shall be compliant with ISO 24613-1, ISO 24613-2, and ISO 24613-3 when implementing data categories referred to in the respective parts. LBX extends the original models by means of data category selections and precise value lists, the creation of new subclasses, and the definition of new constraints. In addition, this document complies with the cardinalities expressed in ISO 24613-1, ISO 24613-2, and ISO 24613-3. The LBX serialization is richer in detail than LMF, in order to meet specific design objectives. Still, this

ISO/DIS 24613-5:2020(E)

document does not elaborate on the meta-data aspects from LMF, since the LBX schema is by essence much richer for the representation of all the aspects related to the creation, content, versioning and database implementation of lexical content at large. Occasionally, slightly equivalent constructs to explicit requirements from the LMF standard will be mentioned.

The XML examples in this document are simplified by omitting namespaces. Except where otherwise stated, it is assumed that XML elements belong to the LBX namespace and that the examples lie within the scope of the following XML namespace declaration:

```
xmlns="http://www.lbx.org/2020/schema"
```

5 Serialization of the LMF core model (ISO 24613-1)

5.1 Implementing the LexicalResource class

The LexicalResource class shall be implemented in LBX by means of the <LexicalResource> element (see Table 1), which groups together one to many lexicons in a single collection. This level may be omitted in cases where the lexical resource contains only one lexicon so that the resource starts directly with the lexicon level. In cases where a lexical resource contains a large number of lexicons or several very large lexicons, the lexicon (XML document) can reference a virtual lexical resource using a @lexicalResourceID in the <Lexicon> element and optionally the <LexicalEntry> element (see 5.5).

Table 1 — LexicalResource class

LMF class	LBX construct
/LexicalResource/	<LexicalResource>

5.2 Implementing the GlobalInformation class

The GlobalInformation class shall be implemented in LBX by means of the <GlobalInformation> element (see Table 2) either by referencing a GlobalInformation.xsd schema using an <xsd:include> element, or as a direct child of a <LexicalResource> element. <GlobalInformation> allows the encoding of a variety of administrative, technical, documentary, and bibliographic information attached to the corresponding lexical resource.

Table 2 — GlobalInformation class

LMF class	LBX construct
/GlobalInformation/	<GlobalInformation>

Since the LBX serialization is based on the W3C recommendation for XML, it implements the @xml:lang attribute to indicate language information corresponding to the content of specific elements. According to the W3C recommendation, @xml:lang content shall be compliant with BCP 47. There is no need for a specific implementation of the /language coding/ data category or the /script coding/ data category in order to ensure compliance of this document with ISO 24613-1. LBX does allow the inclusion of these data categories in the <GlobalInformation> element in order to support the validation of equivalent metadata found in the <LexiconInformation> elements of one or more lexicons (see 5.4). When included, the /script coding/ shall use the codes from ISO 15924. The /character encoding/ data category is implemented in the XML declaration of an LBX conformant document using the @encoding attribute. For instance, an XML-LBX document encoded as UTF-8 according to the Unicode standard shall begin with the following declaration:

```
<?xml version="1.0" encoding="UTF-8" ?>
```

A non-exclusive list of <GlobalInformation> sub-elements, simple types indexed by value, follows:

- “ISO639-3”, a simple type enumerating the set of language codes used across all lexicons;

- “ISO15924”, a simple type enumerating the set of scripts used across all lexicons;
- globalNotationType, a simple type enumerating the set of notations used across all lexicons;
- globalPartOfSpeechType, a simple type enumerating the set of <partOfSpeech> values used across all lexicons;
- subjectFieldType, a simple type enumerating the set of <SubjectField> values used across all lexicons.

Examples can be found in the LBX reference schema, GlobalInformation document (see [Annex B](#)).

5.3 Implementing the Lexicon class

The Lexicon class is implemented in LBX by means of the <Lexicon> element (see [Table 3](#)), which is a direct child of the <LexicalResource> element when <LexicalResource> is used. If the <LexicalResource> element is not used, <Lexicon> becomes the root element. In cases where a lexical resource contains a large number of lexicons or several very large lexicons, the lexicon (XML document) can reference a virtual lexical resource using a @lexicalResourceID in the <Lexicon> element (see [5.1](#)). In the case of a virtual lexical resource, where the <LexicalResource> element is not part of the same XML document as the <Lexicon> element, the lexicon can use an include statement to reference a relevant <GlobalInformation> element. Other information within the <Lexicon> element should be qualified through the following child element(s) and attributes as direct children of the <Lexicon> element or, optimally, as children of the <LexiconInformation> element (see [5.4](#)):

- <Title>, the title of the lexicon;
- @lexiconID, of datatype xs:ID as a unique identifier for the lexicon; as a best practice, the id should be a URI and be unique within a language resource. @xml:ID can be used in place of @lexiconID when there is a design intent to make the entry accessible on the web;
- @lexicalResourceID of datatype xs:ID as a unique identifier for the lexical resource; as a best practice, the ID should be a URI for global scope; in addition, @xml:ID can be used in place of @lexicalResourceID when there is a design intent to make the entry accessible on the web;
- @lexiconType, of @datatype “xs:string”; the type of lexicon, e.g. bilingual dictionary, monolingual dictionary;
- @sourceLanguage, of @datatype “xs:string”; the language of the <Lemma> element or its inflected forms;
- @targetLanguage, of @datatype “xs:String”; the language the Lemma is translated to, principally represented in the <Translation> element.

Table 3 — Lexicon class

LMF class	LBX construct
/Lexicon/	<Lexicon>

5.4 Implementing the LexiconInformation class

The LexiconInformation class is implemented by means of the LBX <LexiconInformation> element (see [Table 4](#)) either by referencing a LexiconInformation.xsd schema using an <xsd:include> element or as a direct child of the <Entry> element. <LexiconInformation> allows the encoding of a variety of administrative, technical, documentary, and bibliographic information attached to the corresponding lexical entry.

Table 4 — LexiconInformation class

LMF class	LBX construct
/LexiconInformation/	<LexiconInformation>

When not included in the <Lexicon> element, information qualifying the lexicon should be included as elements and attributes in the <LexiconInformation> element. These include (see 5.3):

- <Title>;
- @lexiconID
- @lexicalResourceID;
- @lexiconType;
- @sourceLanguage;
- @targetLanguage.

The <LexiconInformation> can also include elements and data categories that further qualify information in the lexicon and can be used to support the validation of the XML document (lexicon). These elements and data categories should also be included in the global set of elements and data categories found in the <GlobalInformation> element (see 5.2) and a comparison of the corresponding values in <GlobalInformation> and <LexiconInformation> should be part of the validation process.

A non-exclusive list of these sub-elements, simple types indexed by value, follows:

- notationType, a simple type enumerating the set of notations used in a lexicon;
- partOfSpeechType, a simple type enumerating the set of <partOfSpeech> values used in a lexicon;
- subjectFieldType, a simple type enumerating the set of <SubjectField> values used in a lexicon.

Examples can be found in the LBX reference schema, LexiconInformation document (see B.1).

NOTE In addition to the <LexiconInformation> construct, LBX allows the concatenation of lexicon information for a subset of lexicons grouped by language by referencing a named language data schema (e.g. ArabicLanguageData.xsd) (see B.1).

5.5 Implementing the LexicalEntry class

The LexicalEntry class should be implemented by means of the <Entry> element in LBX (see Table 5). Lexical information inside <Entry> elements should be encoded through the following child elements:

- <GramFeats> for grammatical information related to the whole entry;
- <Form> for containing the text literal and attributes qualifying the text literal (the Form class is serialized through subclasses in LBX);
- <Etymology> for etymological aspects;
- <Sense> for semantic information;
- <Xref> for referencing internal or external elements.

Attributes used for the <LexicalEntry> element can include:

- @entryID of datatype xs:ID as a unique identifier for an entry; as a best practice, the id should be a URI and be unique within a language resource; @xml:ID can be used in place of @entryID when there is a design intent to make the entry accessible on the web;