INTERNATIONAL STANDARD

ISO 24613-5

First edition

Language resource management — Lexical markup framework (LMF) —

Part 5: Lexical base exchange (LBX) serialization

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

Partie 5: Sérialisation de l'échange de bases lexicales (LBX)

PROOF/ÉPREUVE

Reference number
ISO 24613-5:2021(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

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 24613-1:2019, ISO 24613-2:2020, ISO 24613-3:2021 and ISO 24613-4:2021, cancels and replaces ISO 24613:2008, which has been technically revised.

The main change compared to the previous edition is as follows:

— entire revision of the content and its subdivisions into several parts.

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.

Language resource management — Lexical markup framework (LMF) —

Part 5: Lexical base exchange (LBX) serialization

1 Scope

This document describes the serialization of the lexical markup framework (LMF) model defined as an extensible markup language (XML) model derived from the language base exchange (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.

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*

IETF BCP 47. Tags for Identifying Languages. Phillips, A., Davis, M. (eds.), September 2009. Best Current Practice. Available from: https://tools.ietf.org/html/bcp47

W3C. Extensible Markup Language (XML) 1.1 (Second Edition). W3C Recommendation 16 August 2006, edited in place 29 September 2006. Available from: https://www.w3.org/TR/2006/REC-xml11-20060816/

3 Terms and definitions

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

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 requires compliance with ISO 24613-1, ISO 24613-2, and ISO 24613-3 when implementing data categories referred to in the respective parts, and compliance with the W3C XML Schema 1.1 for representing structured information in XML. 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 document does not elaborate on the metadata 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 are 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.LexicalBaseExchange.org/2021/schema"
```

Besides, datatypes in this document are defined in compliance to the XML Schema Part 2 recommendation. The “xs:” prefix corresponds to the namespace http://www.w3.org/2001/XMLSchema.

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:

- "ISO 639-3", a simple type enumerating the set of language codes used across all lexicons;
- "ISO 15924", 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 shall be 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 shall be implemented in LBX by means of the <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 shall 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.

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 Clause B.1).

5.5 Implementing the LexicalEntry class

The LexicalEntry class shall be implemented in LBX by means of the <Entry> element (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;
- @lexiconID of datatype xs:ID as a unique identifier for the parent 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 @ entryID when there is a design intent to make the lexicon accessible on the web;
- @lexicalResourceID, a reference to the @lexicalResourceID of the associated lexicon collection when there is more than one lexicon.

Table 5 — LexicalEntry class

LMF class	**LBX construct**
/LexicalEntry/	<Entry>

The following example in French illustrates the encoding of a simple dictionary entry with two senses.

EXAMPLE

```
<Entry xml:lang="fr">
    <Etymology>XIIIe; languste, v. 1120, «sauterelle»; encore dans Corneille (Hymnes, 7);
anc. provençal langosta, altér. du lat. class. locusta «sauterelle».</Etymology>
    <Lemma>
         <GramFeats>
             <POS>noun</POS>
             <Gender>fem</Gender>
         </GramFeats>
        <FormRep xml:lang="fr" notation="French">langouste</FormRep>
        <FormRep xml:lang="fr" notation="IPA">lãgust</FormRep>
    </Lemma>

    <Sense senseNR="1">
         <Def>
            <DefRep xml:lang="fr">Grand crustacé marin (Décapodes macroures) aux pattes
antérieures dépourvues de pinces, aux antennes longues et fortes, et dont la chair est
très appréciée.</DefRep>
         </Def>
    </Sense>

    <Sense senseNR="2">
         <Note type="socioCultural">Fig. et fam. (vulg.).</Note>
         <Def>
            <DefRep xml:lang="fr">Femme, maîtresse</DefRep>
         </Def>
    </Sense>
</Entry>
```

NOTE 1 The style in the above example is appropriate for use in a lexical resource that contains a collection of bilingual lexicons in a variety of source languages, e.g. French, Spanish, Russian, Chinese. A simpler style can be used for a collection of monolingual French lexicons. For example, <Orth> and <Pron> can be used in place of the equivalent <FormRep> elements and the <Def> element can directly contain the text content rather than employing a <DefRep> child element for managing text content (see 5.10). See 6.2 for an example of simplification using the <Orth> and <Pron> elements.

NOTE 2 The @notation value “French” is short for “Canonical French”.

5.6 Implementing the OrthographicRepresentation class

Classes containing an OrthographicRepresentation class include the Form, Lemma, and Definition classes. Orthographic representations shall be implemented in LBX by means of elements corresponding to OrthographicRepresentation subclasses that are introduced in ISO 24613-2 (machine-readable dictionary (MRD) model), or possible new OrthographicRepresentation subclasses derived through the principles for LMF extensions described in ISO 24613-1 (Core model). ISO 24613-1:2019, 5.6.1, describes some of the representation types that can serve as a basis for extending the OrthographicRepresentation class. ISO 24613-4 (TEI extension), 6.1, lists a number of representation elements that are valid for use with the Form class. Elements implemented in this part are described in 5.7.2, 5.10, and successive subclauses from 6.3.2 to 6.3.8.

5.7 Implementing the Form class

5.7.1 Form class

The Form class shall be implemented in LBX by means of Form subclasses (see Table 6, 6.2 and 6.3).

Table 6 — Form class

LMF class	LBX construct
/Form/	<Form>

5.7.2 Lemma class

The Lemma class, a subclass of the Form class, shall be implemented in LBX by means of the <Lemma> element (see Table 7).

Table 7 — Lemma class

LMF class	LBX construct
/Lemma/	<Lemma>

Orthographic representations in the <Lemma> element shall be implemented in LBX by means of the <FormRep> element, or by elements that instantiate <Form> subclasses, including <Orth> and <Pron>.

NOTE 1 The <FormRep>, <Orth>, and <Pron> elements are introduced in 6.2.

NOTE 2 <Orth> and <Pron> can be allowed when justified by design goals.

5.8 Implementing the GrammaticalInformation class

The GrammaticalInformation class groups grammatical features associated with the LexicalEntry class, Form class, or other classes (e.g. Translation, Sense) in case of specific grammatical restrictions. The GrammaticalInformation class shall be implemented in LBX by means of the <GramFeats> element (see Table 8) combined with various possible child elements for specific grammatical features.

Table 8 — GrammaticalInformation class

LMF class	LBX construct
/GrammaticalInformation/	<GramFeats>

LBX provides the following child elements of <GramFeats> for describing specific grammatical features of associated elements (e.g. <Lemma>, <WordForm>):

— <POS> to indicate the grammatical category of the lexical item. This corresponds to the /partOfSpeech/ data category in ISO 24611:2012, Annex A;

— <Person> to indicate the grammatical person (if relevant) of the lexical item or one of its inflected forms. This corresponds to the /person/ data category in ISO 24611:2012, Annex A;

— <Gender> to indicate the grammatical gender (if relevant) of the lexical item or one of its inflected forms. This corresponds to the /grammaticalGender/ data category in ISO 24611:2012, Annex A;

- <Number> to indicate the grammatical number (if relevant) of the lexical item or one of its inflected forms. This corresponds to the /grammaticalNumber/ data category in ISO 24611:2012, Annex A;
- <Tense> to indicate the grammatical tense (if relevant) of the lexical item or one of its inflected forms. This corresponds to the /grammaticalTense/ data category in ISO 24611:2012, Annex A;
- <Aspect> to indicate the grammatical aspect (if relevant) of the lexical item or one of its inflected forms;
- <Mood> to indicate the grammatical mood (if relevant) of the lexical item or one of its inflected forms;
- <Voice> to indicate the grammatical voice (if relevant) of the lexical item or one of its inflected forms;
- <Animacy> to indicate the grammatical animacy (if relevant) of the lexical item or one of its inflected forms (e.g. in Russian);
- <GrammaticalClass> to indicate the grammatical class (gender) of Bantu languages;
- <GrammaticalClassGroup> to indicate the aggregate grammatical classes (genders) of a specific noun in the singular and plural;
- <iType> to indicate the inflectional class associated with the lexical item or one of its inflected forms;
- <Subcat> to indicate subcategorization information (e.g. transitive/intransitive, countable/non-countable).

The following example shows the grammatical information for a word form in a monolingual French dictionary that is part of a notional language resource containing a collection of monolingual and bilingual dictionaries in multiple source languages. The @notation="French", denoting canonical French, is used in databases that support a large set of possibly idiosyncratic notations (e.g. for canonical, transliterated and transcribed forms).

EXAMPLE

```
<Entry>
     <Lemma>
         <GramFeats>
             <POS>verb</POS>
             <Subcat>transitive</Subcat>
         </GramFeats>
         <FormRep xml:lang="fr" notation="French">pacifier</FormRep>
         <FormRep xml:lang="fr" notation="ipa">pasifje</FormRep>
      </Lemma>
      <Sense></Sense>
</Entry>
```

For an example of simplifying this schema, see 6.2.

5.9 Implementing the Sense class

The Sense class, as a recursive construct, shall be implemented in LBX by means of the <Sense> element (see Table 9). LBX does not allow character content in the element.

Table 9 — Sense class

LMF class	LBX construct
/Sense/	<Sense>