DRAFT INTERNATIONAL STANDARD

ISO/DIS 24613-3

ISO/TC **37**/SC **4**

Secretariat: **KATS**

Voting begins on:
2020-02-17

Voting terminates on:
2020-05-11

Language resource management — Lexical markup framework (LMF) —

Part 3:
Etymological extension

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

Partie 3: Extension étymologique

ICS: 01.020

This document is circulated as received from the committee secretariat.

Reference number
ISO/DIS 24613-3:2020(E)

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

Contents

Page

Foreword ..... **iv**

1 Scope **1**

2 Normative references **1**

3 Terms and definitions **1**

4 The LMF Etymology extension **2**
4.1 The Cognate class and the Etymon class 2
4.2 The Etymologisable class 2
4.3 The Etymology class and the EtyLink class 2
4.4 The CognateSet class 3
4.5 The Date class 4

Annex A (informative) **Examples of possible etymological typologies** **5**

Annex B (normative) **Data categories for etymology description** **13**

Bibliography **19**

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 24613-3, together with ISO 24613-1, ISO 24613-2, ISO 24613-4, and ISO 24613-5, 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:

— entire revision of the content and its subdivision.

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 3: Etymological extension

1 Scope

This document describes an extension to ISO 24613-1 and ISO 24613-2 to support the development of detailed descriptions of common etymological phenomena and/or diachronic information with respect to lexical entries in born-digital and/or retro-digitized lexicons. It provides both a meta-model for such an extension as well as the relevant data categories.

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*

ISO 24613-2, *Language resource management — Lexical Markup Framework (LMF) — Part 2: Machine Readable Dictionary (MRD) model*

3 Terms and definitions

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

3.1
etymologisable
capable of having an etymology

Note 1 to entry: "Etymologisable" is a category of lexical elements and usages (encompassing for instance lexical entries, senses, word forms).

3.2
etymology
origin and historical development of any aspect of a given lexical item

3.3
etymon
lexical entry from which another lexical entry is derived

Note 1 to entry: An etymon may also be simply an earlier stage of a lexical item.

3.4
cognate
form in a related language which has been derived from the same etymological origin as another form

4 The LMF Etymology extension

4.1 The Cognate class and the Etymon class

Cognate and Etymon are defined as subclasses of the LexicalEntry class from the LMF core module (see Figure 1)[1]). Both classes define lexical entries whose purpose in a lexical resource is to help describe the etymology of one or more lexical entries. Note that instances of either Etymon or Cognate can be assigned a language different from the language of the lexicon as a whole (recall that this is specified in the LexiconInformation class in ISO 24613-1).

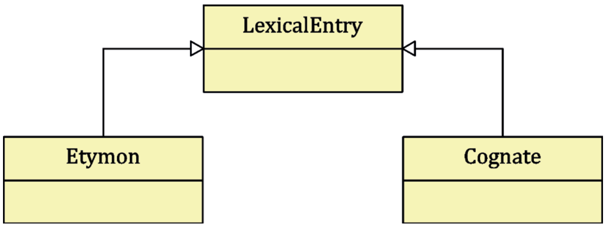

Figure 1 — Cognate and Etymon as subclasses of LexicalEntry

Individuals of both the Etymon and the Cognate class shall be in an aggregation relationship with at least one individual of type EtyLink (see 4.3). Note that when describing etymologies it can be necessary to deal with instances of LexicalEntry (and hence also instances of Etymon and Cognate) which are roots, and especially reconstructed roots. In this case the type of the root in question can be specified using the attribute *@rootType*. In the case of reconstructed roots or other word forms the attribute @ *status* serves to associate the element with a verbal description of the likelihood of its having been in use (see the example in A.8).

4.2 The Etymologisable class

The Etymologisable class provides a means of referring to the set of linguistic elements that can have etymologies. By defining a single class encompassing all such 'etymologisable' elements, the types of elements which can have etymologies may be easily extended in the future as the necessity arises. The following classes are subtypes of the Etymologisable class (see Figure 2): LexicalEntry, Sense, Form and CognateSet (see 4.4).

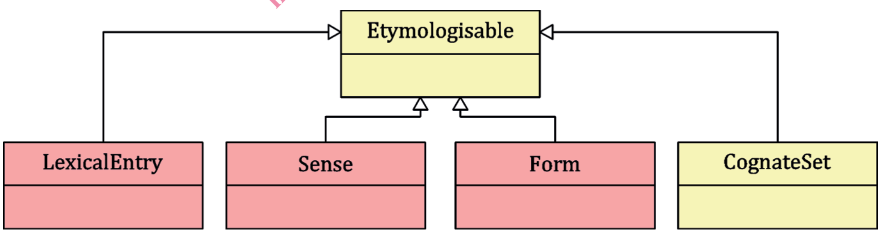

Figure 2 — The Etymologisable class and its subclasses

4.3 The Etymology class and the EtyLink class

The Etymology class allows for the description of the etymology of a linguistic element, i.e., those linguistic elements which are subclasses of the Etymologisable class. A typology of etymological processes may be specified for a given Etymology using the *@type*, and also potentially the *@subtype* attribute (in the case when the type of the etymology can be further specified), possible values may vary according to the theoretical approach adopted and/or linguistic or editorial focus of the resource in question. Note that by using nested Etymology instances, a combination of processes is permitted.

1) Note that in this document, the following colour scheme is used in diagrams: classes in yellow are introduced in this document, and classes in pink have been previously introduced in ISO 24613-1 and ISO 24613-2.

Examples of such etymological processes are: *borrowing, inheritance; word-formation: compounding, derivation; sense shifts: narrowing, widening, amelioration, pejoration, metaphor, metonymy; phonetic/phonological processes: place assimilation, dissimilation, epenthesis, metathesis, hardening, weakening, etc.* (these data categories are thoroughly described in Annex B). Individual links between two elements in an etymology may also be given a type, see the description of EtyLink below. Given that an Etymology instance may be taken from an external source, it may be associated with a Bibliography instance, as defined by ISO 24613-2 (see Figure 3).

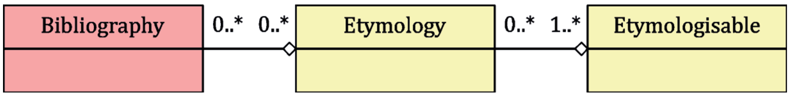

Figure 3 — The Etymology class

Instances of Etymology are associated with one or more EtyLink instances, each of which represents a single stage or step in the etymology of a given lexical item (see Figure 4). EtyLink serves to link together individuals belonging to the subclasses of Etymologisable. EtyLink is a subclass of the CrossREF class defined in ISO 24613-1. The use of CrossREF requires that the target objects representing the given lexical content be given *@id* attributes. By making use of the *@id* attribute on an individual of the Etymologisable class as a target, a generic sequential temporal ordering of multiple elements can be modelled using the attributes *@prev* and *@next*. Instances of the EtyLink class can further specify other temporal relationships using various temporal attributes associated with the source and target of each EtyLink instance.

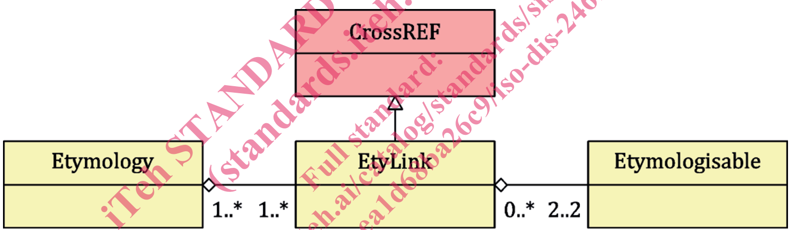

Figure 4 — EtyLink

Note that individuals of the Etymon and Cognate classes (subtypes of LexicalEntry) shall be associated with at least one individual of EtyLink.

4.4 The CognateSet class

The CognateSet class (see Figure 5) is a container for sets of one or more Cognate items and zero or more Bibliography items (see ISO 24613-2). The CognateSet is an onomasiological construct, the contents of which are items from related languages to that of the Etymon and/or LexicalEntry that are gathered with the purpose of demonstrating linguistic similarities or dissimilarities of some kind. A CognateSet inherently implies that the Cognate instances and the LexicalEntry share an etymological source. It can be linked directly to LexicalEntry or it can occur as a subclass of Etymology.

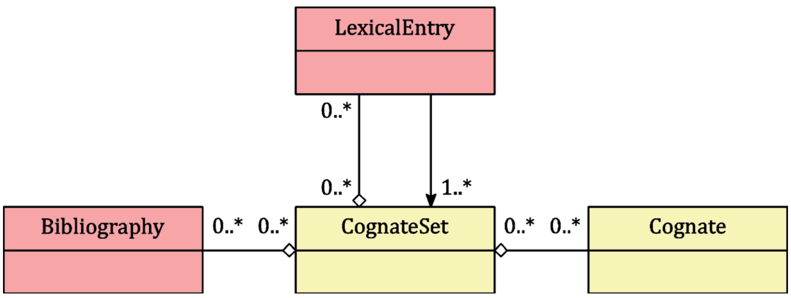

Figure 5 — CognateSet

4.5 The Date class

Components of a LexicalEntry and its subclasses can be associated with a specific date by making use of the Date class. Furthermore Date can be used to specify a number of degrees of precision. A precise year, or potentially day and/or month, can be stated using the @*date* attribute and a rough date with the attribute @*circa*. Within a span of time with different levels of specificity there is the possibility of using one or more dating attributes. Where a span of time is known (or asserted) the lower and upper ends of the span can be specified using @*notBefore*, @*notAfter* respectively.

Table 1 — Example of class adornment

Class Name	Example of Attributes	Comment
Etymon	xml:lang, gloss	
Etymology	type, subtype	
EtyLink	type, prev, next	
CognateSet		
Cognate	xml:lang, gloss	
Date	notBefore, notAfter, circa, date	

Annex A
(informative)

Examples of possible etymological typologies

A.1 Example of simple inheritance

The example in Figure A.1 describes the inheritance of a lexical entry from a parent language, in this case the adverb *semper* in Sardinian which comes from the Latin word *semper*. The Sardinian word is linked to a single Etymology instance which is associated with the type *inheritance* (see Annex B for a definition of this type). The Etymology is then associated with an individual of type EtyLink which represents the process of change from the Latin etymon to the Sardinian lexical entry.

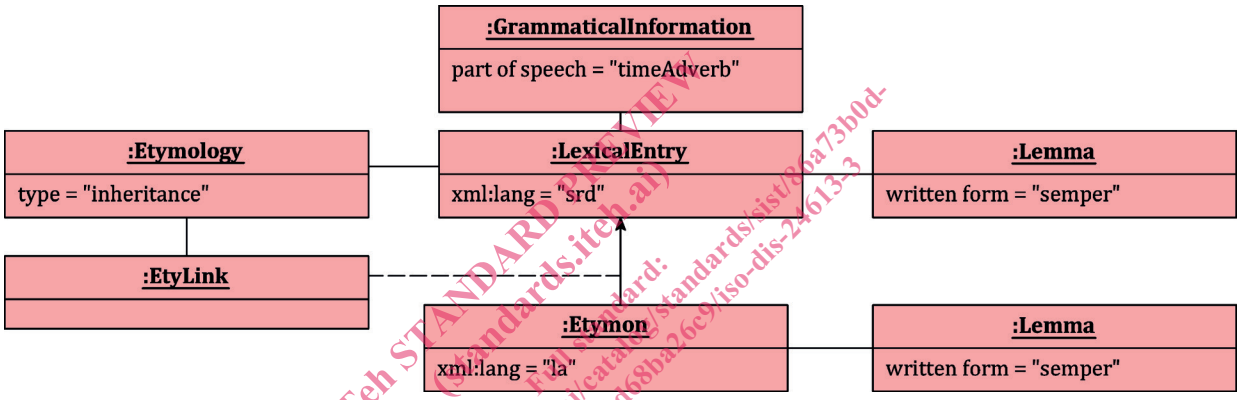

Figure A.1 — Diagram of inheritance in Sardinian

A.2 Example of a diachronic etymological process (inheritance with phonological change)

In the following example the development of the word meaning ‘wine’ (ipa: [veŋ]) in the Emiliano variety of Italian is traced using a series of Etymons which are ordered and linked together using EtyLink instances. These take us from the Vulgar Latin *vinu* through to the immediate predecessor of the word in its current manifestation. Once again these individual links can be accessed through an Etymology individual (with the type *inheritance*) which represents the history of the LexicalEntry. In Figure A.2 there is no explicit ordering of the EtyLinks, but this can be added using the attributes @*prev* and @*next*. (These have not been added to the diagram for reasons of space, but they have been implemented in the example in A.6.)

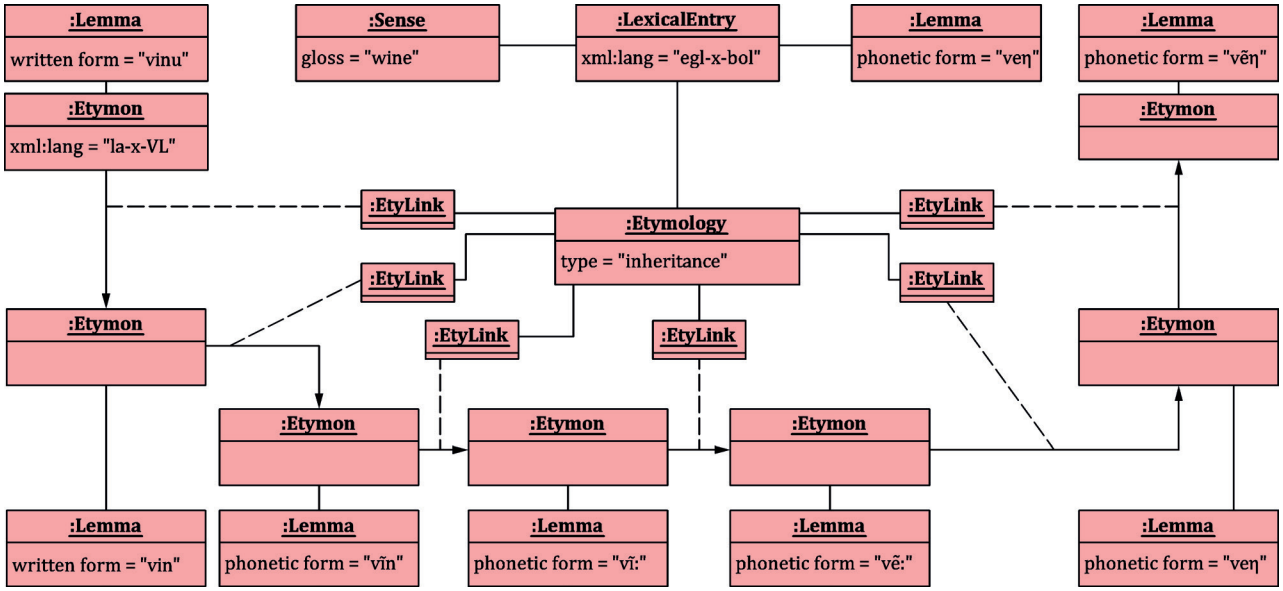

Figure A.2 — Diagram of multi-stage inheritance and phonological change in Bolognese

A.3 Example of word form inheritance

The derivation of the singular and plural forms of a Portuguese noun *naçao* (sg), *nações* (pl), respectively, derived from two forms of a Vulgar Latin (VL) noun, *nātiōnem* (sg, acc), *nātiōnes* (pl, acc) is described in Figure A.3. Herein, since the Portuguese forms concerned are both the plural and singular, the Etymon has two WordForm instances, one for each grammatical number. Note in particular the association of *@grammaticalCase* and *@inflectionType* attributes with the WordForm of the Etymon via a GrammaticalInformation instance. In a comprehensive lexicon of Portuguese that contained such etymological information for a sufficient number of lexical entries, it would be possible, by contrasting the contents of the WordForm in the LexicalEntry with the Etymon, to appreciate the following language-wide phenomena: 1) Portuguese lost grammatical case; 2) the vast majority of its nouns came from the VL accusative case; 3) where the VL singular (accusative) ending was *-tiōnem*, the Portuguese form is written "-çao" and pronounced [sɐ̃ w]; where the VL plural (accusative) ending was *-tiōnes*, the Portuguese form is written "-ções" and pronounced [sõj̃ s].

NOTE This etymology could be further articulated by adding the phonological process types for each stage of the diachrony. This would be done in the model by adding the appropriate data category defined in Annex B to the value of *@type* on the EtyLink for the given stages.