

Modelling Etymology in LMF/TEI: The Grande Dicionário Houaiss da Língua Portuguesa Dictionary as a Use Case

Fahad Khan, Laurent Romary, Ana Salgado, Jack Bowers, Mohamed Khemakhem, Toma Tasovac

▶ **To cite this version:**

Fahad Khan, Laurent Romary, Ana Salgado, Jack Bowers, Mohamed Khemakhem, et al.. Modelling Etymology in LMF/TEI: The Grande Dicionário Houaiss da Língua Portuguesa Dictionary as a Use Case. LREC 2020 - 12th Language Resources and Evaluation Conference, May 2020, Marseille, France. hal-02618067

HAL Id: hal-02618067
https://inria.hal.science/hal-02618067

Submitted on 25 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Modelling Etymology in LMF/TEI: The *Grande Dicionário Houaiss da Língua Portuguesa* Dictionary as a Use Case

Fahad Khan[1], **Laurent Romary**[2], **Ana Salgado**[3,4], **Jack Bowers**[2,5], **Mohamed Khemakhem**[6,2,7,8], **Toma Tasovac**[9]
[1]Istituto di Linguistica Computazionale "A. Zampolli– CNR", Pisa, Italy
[2]Inria-ALMAnaCH – Automatic Language Modelling and ANAlysis Computational Humanities, Paris, France
[3]NOVA CLUNL, Universidade NOVA de Lisboa, Lisbon, Portugal
[4]Academia das Ciências de Lisboa, Lisbon, Portugal
[5]ACDH-CH – Austrian Center for Digital Humanities and Cultural Heritage, Vienna, Austria
[6]Litt & Arts - UMR 5316, Grenoble
[7]Université Paris Diderot, Paris, France
[8]Centre Marc Bloch, Berlin, Germany
[9]Belgrade Center for Digital Humanities, Belgrade, Serbia
fahad.khan@ilc.cnr.it, {laurent.romary, mohamed.khemakhem}@inria.fr,
anasalgado@campus.fcsh.unl.pt, Jack.Bowers@oeaw.ac.at, ttasovac@humanistika.org

Abstract

In this article, we will introduce two of the new parts of the new multi-part version of the Lexical Markup Framework (LMF) ISO standard, namely Part 3 of the standard (ISO 24613-3), which deals with etymological and diachronic data, and Part 4 (ISO 24613-4), which consists of a TEI serialisation of all of the prior parts of the model. We will demonstrate the use of both standards by describing the LMF encoding of a small number of examples taken from a sample conversion of the reference Portuguese dictionary *Grande Dicionário Houaiss da Língua Portuguesa*, part of a broader experiment comprising the analysis of different, heterogeneously encoded, Portuguese lexical resources. We present the examples in the Unified Modelling Language (UML) and also in a couple of cases in TEI.

Keywords: LMF, TEI, Portuguese Language Resources, Dictionaries

1. Introduction

In this article, we will introduce two parts of the new multi-part version of the Lexical Markup Framework (LMF) International Organization for Standardization (ISO) standard, namely Part 3 of the standard (ISO/DIS 24613-3), which deals with etymological and diachronic data, and Part 4 (ISO/DIS 24613-4), which consists of a TEI serialisation of all of the prior parts of the model (in what follows we will refer to the TEI-XML serialisation of LMF data as *LMF in TEI*). In particular, we will show how LMF and especially parts 3 and 4 can be used to encode etymological data by taking example encodings of entries from an important reference dictionary for the Portuguese language, the *Grande Dicionário Houaiss da Língua Portuguesa* (Houaiss, 2015), from now on *Houaiss*. Note that although we have previously introduced the new version LMF standard as a whole (Romary et al., 2019), our aim in the current work is to describe and motivate the particular approach which we, as authors of these two parts of the standard, have taken to representing etymological data, an approach which is closely related to previous work carried out by some of the authors of the current work in TEI and linked data, but which also takes advantage of the increased levels of abstraction offered by LMF over those two other data frameworks. The rest of the article is as follows. In Section 2., we discuss the LMF as a whole in order to situate ISO 24613-3 and ISO 24613-4 within the broader context of the entire standard; the latter two parts are described in detail in Sections 2.1. and 2.2. respectively. In Section 3., we will give some brief background on the resource (*Houaiss*) which will furnish us with almost all of our examples in this paper. Next, in Section 4., we will give some illustrative examples from the Houaiss dictionary and describe their encoding in LMF using UML; we will describe the TEI encoding of two of these examples in Section 5. In the conclusion, we describe future work.[1]. The appendix includes further details on combining etymological information from different resources.

2. The Lexical Markup Framework

The Lexical Markup Framework was first published as a standard in 2008 by ISO (ISO 24613: 2008) having been intended from the first as a "standard framework for the construction of computer lexicons" (Francopoulo et al., 2006). Since that time, it has been used in the creation and publication of lexical resources by several organizations and numerous national and international projects (Francopoulo, 2013). However, as

[1]It should be noted that the present work is not intended to replace the contents of the standard being described. Instead, it is a description of a specific use case for the standard.

a result of a detailed review of the standard, it was decided, primarily for reasons of lack of modularity and the inadequacy of the serialisation initially associated to the model, that the ISO working group should revise LMF. ISO TC 37/SC 4/WG 4 (*Lexical resources*) and become a multi-part standard. At the time of writing the first part of the new version has already been published[2], and four of the other parts are at an advanced stage of completion. The first two parts of the new standard deal with aspects of linguistic description that were already covered in the previous version of LMF but the third and fourth parts are new; the former provides a data model for publishing etymological and (more generally) diachronic lexical information, and the latter constitutes an XML serialisation of the preceding parts following the Text Encoding Initiative guidelines. In summary, the different parts of the revised standard are as follows[3]:

- **Core Model (ISO 24613-1:2019)**, **Machine Readable Dictionaries (ISO 24613-2)**: described below in Section 2.1.; **Diachrony-Etymology (ISO 24613-3)**: described below in Section 2.2.; **TEI serialisation (ISO 24613-4)**: described below in Section 2.3.;
- **LBX serialisation (ISO 24613-5)**: a second serialisation is formalised here using the Language Base Exchange (LBX); **Syntax and Semantics (ISO 24613-6)**: this part gathers together classes and attributes necessary for a detailed encoding of semantic and syntactic information; **Morphology (ISO 24613-7)**: this part gathers together classes and attributes necessary for a detailed encoding of semantic and syntactic information – we do not discuss these parts here.

2.1. LMF Parts 1 (ISO 24613-1:2019) and 2 (ISO 24613-2)

The LMF core model, as described in Part 1 (ISO 24613-1:2019) of the standard, consists of classes and attributes deemed to be fundamental for the modelling of computational lexica. It still remains very close to the definition of the core properties of the original 2008 version of LMF. Note also that unlike other parts of the standard its use is mandatory in the definition of LMF lexicons. The LMF core model includes the class *Lexical Resource*, which serves to group together different individuals of the class *Lexicon*; the latter class standing as a container for a collection of individuals of the class *Lexical Entry*. Other fundamental core model classes include *Form*, *Sense*, and *Definition*. One crucial novelty of Part 1 is the class *CrossRef*, which was not available in the 2008 version of LMF, and which provides a generic linking mechanism for the other classes in the standard: both in the case of linking to external as well as internal resources. Most importantly for our purposes, and as we shall see in the next section, *CrossRef* is used in the modelling of etymological links. Part 2 of LMF, the Machine Readable Dictionaries module (ISO 24613-2), also includes a number of classes which are essential for our purposes in describing ISO 24613-3. These include *Word Form* and *Bibliography*.

2.2. LMF Part 3 (ISO/DIS 24613-3)

Our approach to modelling etymologies in LMF has been heavily influenced by prior work in etymology representation both in the 2008 version of LMF (Salmon-Alt, 2006) as well as in other standards such as TEI (Bowers and Romary, 2017) and Ontolex-lemon (Khan, 2018). In particular, we have sought to emphasise the following three aspects of etymologies when representing them as computational resources:

1. their **status as descriptions of abstract graphs**, i.e., data structures of the kind that computer scientists are (exceedingly) used to working with;
2. the fact that **etymologies usually describe the genealogy of a given lexical phenomenon through the use of a narrative** (simple in most cases); and finally,
3. the fact that individual **etymologies also describe hypotheses** (often of a scholarly nature) and tend to include references to other texts in justification of these hypotheses.

Our more general aim in Part 3 has been to abstract away from individual, written, representations of etymologies and to make etymological data encoded in LMF (and which has been potentially taken from a number of diverse sources, and not just necessarily lexicons) as interoperable as possible: that is we are focused more on the content of etymological information than in the exact form in which it is presented in a source text, as is the case in some TEI based approaches. In order to better motivate our particular approach and to make it more intuitive, we will model a very simple case (abstracted from a class of real examples), passing through different levels of complexity of description – and justifying each successive layer of additional descriptive expressivity – before finally arriving at the approach which has been established in LMF Part 3. This will also allow us to introduce individual classes from the standard in turn and give some justification for their inclusion.
Our example case is one in which we are to model the fact that a word has four etymons, $e1$, $e2$, $e3$, $e4$, in that order (e.g., so that $e4$ derives from or is borrowed from $e3$, $e3$ is derived from or is borrowed from $e2$, and $e2$ comes from $e1$). The first class which we introduce from Part 3, then, is *Etymon* (which we have modelled as a subclass of the class *Lexical Entry* from Part 1 of LMF). A simple way of representing the case in hand would, therefore, be as follows: each instance of *Etymon* links to its succeeding *Etymon* or *Lexical Entry* in the chain as in Figure 1. In fact, this would be sufficient

[2] `https://www.iso.org/standard/68516.html`

[3] Note that LMF is a UML-native framework so that in what follows we will present our LMF examples as Unified Modelling Language (UML) diagrams.

for any such similar case in which an etymology is essentially a single ordered sequence of etymons. The limitations of such an approach are however obvious. It does not allow us to describe the steps between an etymon and its succeeding etymon/entry (for instance, to specify whether one etymon or entry has been been borrowed or inherited from another) in a structured way.

We decided to resolve this issue through the expedient of reifying individual etymological links and introducing the class ***EtyLink*** in Part 3 as a subclass of the *CrossREF* class introduced in Part 1 (where the latter, as we previously mentioned, represents a generic means of linking together elements in an LMF resource). This allows us to associate individuals of type *EtyLink* with each of the etymological links between etymons and etymons/lexical entries in an etymology, as in Figure 2 below. In general, *EtyLink* allows us to predicate any kind of information we choose to of these individual etymological links, including provenance, certainty, etc. However, once more, this does not suffice in cases where we wish to predicate salient properties of the etymology as a whole, such as, again, its provenance, or to assign a level of certainty to the whole thing; neither does it make allowance for cases where a single entry has more than one etymology.

In order to handle these cases (which become more frequent as the resources to be modelled become more 'scholarly') we decided to make etymology a class, ***Etymology***, in Part 3. However since etymologies usually represent an ordering of etymons (and, of course, etymons can be associated with more than one etymology and even more than one etymology for the same entry), we opted to create indirect rather than direct associations between etymologies and etymons. That is, we define etymologies as containers for an ordered series of etymological links, see Figure 3 below. As well as giving us the flexibility we need to describe more than one etymology for the same lexical entry this approach also permits us to re-use etymons and etymological links when relevant, as in Figure 4 where a lexical entry has two etymologies which share three etymons in common.

Summing up, we have so far introduced three of the most important new classes which appear in Part 3: *Etymology*, *Etymon*, and *EtyLink*. Other important classes include *Cognate* and *Cognate Set*, which we will introduce below in the examples in Section 4.

2.3. LMF Part 4 (ISO/DIS 24613-4)

The fourth module in LMF incorporates the encoding alternatives from the TEI standard (Budin et al., 2012) in replacing the old ad-hoc XML serialisation of LMF. Our approach was motivated by earlier work (Romary, 2010; Romary, 2015) which identified modelling analogies between both standards and made concrete proposals towards finding a common serialisation. However given the wide-ranging flexibility that exists in the TEI schema for encoding a given lexical structure, we decided to build on the most common practices within the TEI community while aligning our directives with those of related initiatives, namely the TEI-Lex0 guidelines (Romary and Tasovac, 2018). Part 4 of the standard, therefore provides guidelines for serialising the first three parts and their mutual relationships. In this section, we focus on the serialisation mechanisms including those pertaining to the etymological components of the standard (e.g. Part 3). Following the earlier proposals made in (Bowers and Romary, 2017), the design of a TEI serialisation for Part 3 is based on the following elementary constructs:

- The Etymology class is (obviously) serialised by means of the `<etym/>` element. In case there is a necessity to specify an etymological process it can be expressed by means of the *@type* attribute, whose possible values can be taken from the normative Annex B of LMF Part 3 (e.g. "borrowing", "inheritance", "metaphor", "metonymy", "compounding", "grammaticalization", among others.);

- The Etymon class is interpreted as the citation of a lexical object by means of the `<cit type="etymon"/>` construct which in turn will contain the various components that may characterize a given etymon;

- Referencing a form for an etymon is seen in LMF Part 4 as a descriptive mechanism that parallels the provision of a form in a lexical entry. It is thus implemented with a `<form>` element, together with its possible sub-components (e.g. `<orth>` for orthography or `<pron>` for pronunciation):

```
<form>
  <pron>...</pron>
  <orth>...</orth>
</form>
```

- The CognateSet and Cognate classes are also both serialised by means of the `<cit/>` element with the differentiation being made by specifying the *@type* attribute and the possibility to nest Cognates in a CognateSet construct:

```
<cit type="cognateSet">
  <cit type="cognate">..</cit>
  <cit type="cognate">..</cit>
</cit>
```

- These core constructs are complemented with a variety of components taken up from the TEI guidelines to elicit the language of the etymon (`<lang>`), to express date information (`<date>`), to provide the source of an etymology or an etymon (`<bibl>`).

3. Modelling Etymology in TEI via LMF. A Case Study: the *Grande Dicionário Houaiss da Língua Portuguesa*

In the following sections, we shall illustrate the use of the classes introduced in Section 2.2. by reference to

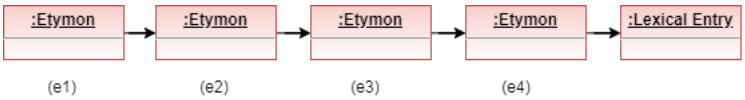


Figure 1: Simple Structured Representation of an Etymon Chain

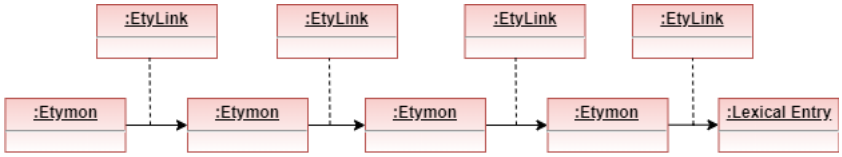


Figure 2: Introducing *EtyLink*

an actual case study: namely, the conversion a number of sample entries from an existing reference dictionary into TEI. The dictionary in question, *Houaiss* (Houaiss, 2015), constitutes an adaptation to European Portuguese of the original Brazilian Portuguese dictionary of the same name (Houaiss, 2001); the latter was the result of the most extensive and ambitious lexicographic project ever realised in Brazil, and was originally published in Rio de Janeiro by the Antônio Houaiss Institute. The adaptation of the dictionary to European Portuguese was carried out by the Dictionary Department of Porto Editora. The resulting edition, published in 6 volumes, was a joint-publication of Círculo de Leitores and the Antônio Houaiss Institute with financial support from the Calouste Gulbenkian Foundation. In Figure 5, we give an example entry from the *Houaiss*. A golden diamond symbol introduces the etymology field of an entry and typically includes an 'immediate' etymon for the entry (the most recent of its etymons listed in the dictionary) and also in many cases other, earlier 'remote', etymons. The immediate etymon is in presented italics, preceded by a source language identifier, and succeeded by a gloss in Portuguese.

4. Example Encodings in LMF

In this section, we will look at three examples from our ongoing conversion of a sample of the *Houaiss* into TEI via LMF[4] and show how they can be represented in LMF UML. The examples, which show the use of the classes introduced in Section 2 in modelling actual dictionary entries, were chosen because of their representativity and in particular, they exemplify the following salient cases: 1) **an undefined etymology** (*cócoras*); 2) **a borrowed lexical unit** (*opalina*); 3) **the existence of a cognate set** (*sapato*).

Our first example is the *Houaiss* entry for the word *cócoras* 'squat', as shown in Figure 6. The etymological section of the *Houaiss* entry for *cócoras* starts with an acknowledgement of the obscurity of the etymology of the word ('etim orig.contrv. ou mesmo obsc.') and follows it up with three different etymological hypotheses. The first, attributed to the Spanish lexicographer Vicente Garcia de Diego[5], proposes an onomatopoeic origin for the word; the second associates it with the Latin etymon *glocire*[6]; the third with the Latin etymon *quacquara*[7]. Our LMF encoding, given in Figure 7, represents each of these etymologies as separate individuals of class *Etymology*, each of which is associated with the *Lexical Entry* for the word. In turn, each of these *Etymology* individuals is associated with an individual of the *EtyLink* class which, in both cases, links the lexical entry with a Latin etymon. The purple coloured boxes in Fig 7 represent the information in the first etymological hypotheses; the light blue boxes the second hypothesis; the light green boxes represent the third hypothesis. Instances of the class *Bibliography* are used to give provenance information for each etymology.

Our second example concerns the Houaiss entry for the word *opalina* 'opaline', see Fig 5, which has three senses: 1) ***a kind of milky translucent glass***; 2) ***objects made from this glass*** (metonomy); 3) ***a white glass used to tile walls, roofs***. The entry's etymology gives the French *opaline* as an immediate etymon; it also specifies that the original sense, not listed as a separate sense in the entry, was 'a parasitic infusory found in frog bellies'[8]. In the LMF encoding, given in Figure

[4]Note that our focus is only on the LMF encoding of the etymological field in each entry. We do not therefore explicitly encode the rest of the information included in the entry here although LMF does offer provision for this across its different modules.

[5]*Garcia de Diego declara-a de orig. onom* ("Garcia de Diego declares it to be of onomatopoeic origin".)

[6]*no REW (no 3795), vincula-se a loc. (ao lado de diversas formas românicas e do v. acocorinhar-se) ao v. lat. glocíre 'cacarejar', prov. por alusão ao choco das galinhas.* ("in REW (no. 3795) it is linked (along with various Romance forms and the verb *acocorinhar-se*) to the vulgar Latin *glocíre* 'cluck' probably by reference to incubation by hens.")

[7]*J. Piel prende, com dúvida, ao lat. popular quacquara 'codorniz', pois, segundo esse autor, teria sido 'a imagem da codorniz agachada no solo (atitude que a confunde com o terreno, tornando-a quase invisível ao caçador), que deu origem ao v. acocorar(-se), acocorinhar(-se), pôr de cócoras etc.'*("J. Piel proposes, although doubtfully, the Vulgar Latin *quacquara* 'quail' because according to this author, 'the image of the quail crouched on the ground (a stance that confounds it with its terrain, making it almost invisible to hunters), may have given rise to the verb *acocorar(-se)*, *acocorinhar(-se)*, *pôr de cócoras* etc")

[8]*a datação é para a acp., não registada aqui, 'género de in-*

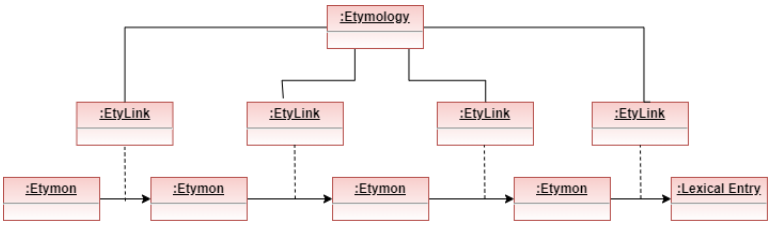


Figure 3: Introducing *Etymology*

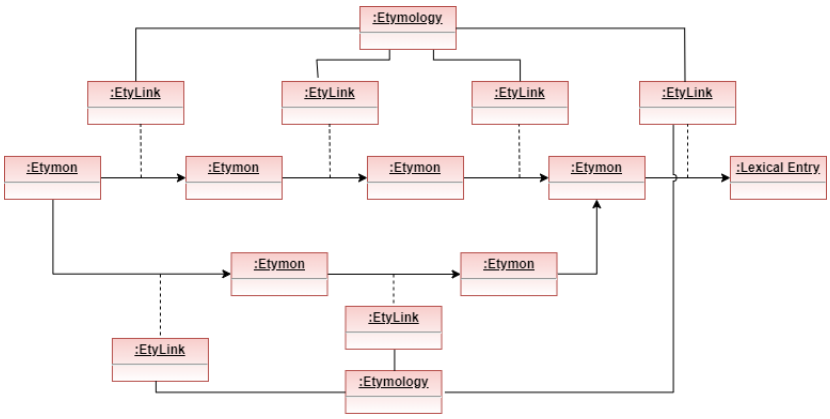


Figure 4: Multiple Etymologies

opalina *s.f.* (1899 cf. CF[1]) 1 vidro de aspeto acetinado, irisado quando visto à contraluz, leitoso, ger. translúcido, us. no fabrico de objetos decorativos, jarras, taças, lustres etc. 2 *p.met.* objeto fabricado com esse vidro ⟨*tem uma coleção de o. francesas*⟩ 3 vidro branco, leitoso, espesso, us. para forrar paredes, tetos etc. ✦ ETIM fr. *opaline* (1899) 'tipo de vaso vitrificado'; ver *opal(i)*-; a datação é para a ace., não registada aqui, género de infusórios que se encontra no ventre das rãs ✦ HOM *opalina* (f. opalino [adj.])

Figure 5: The *Houaiss* entry for *opalina*

cócoras *s.f.pl.* (1561 cf. Lendas) posição agachada → f. menos us.: *cocorinhas* ✦ **a** ou **de** ou **em cócoras** sentado ou apoiado sobre os calcanhares; agachado ✦ USO empr. apenas nestas loc. ✦ ETIM orig.contrv. ou mesmo obsc.; JM limita-se a citar algumas hipóteses: Garcia de Diego declara-a de orig. onom.; no REW (n.° 3795), vincula-se a loc. (ao lado de diversas formas românicas e do v. *acocorinhar*-se) ao v.lat. *glocire* 'cacarejar', prov. por alusão ao choco das galinhas; J. Piel prende, com dúvida, ao lat. popular *quacquara* 'codorniz', pois, segundo esse autor, teria sido «a imagem da codorniz agachada no solo (atitude que a confunde com o terreno, tornando-a quase invisível ao caçador), que deu origem ao v. *acocorar*(-se), *acocorinhar*(-se), pôr de *cócoras* etc.»; a verdade é que não está esclarecida a orig. dessa loc.; f.hist. 1561 *cocoras* ✦ SIN/VAR cócaras

Figure 6: The *Houaiss* entry for *cócoras*.

1, an *Etymology* individual for the entry is associated with an etymological link between a French *Etymon* and the Portuguese *Lexical Entry*[9]. The example also shows an instance of the *Date* class as well as showing how sense glosses are represented using *Definition*.

The next example demonstrates the use of the LMF Part 3 class *Cognate Set*, defined as a container for a set of individuals of the class *Cognate*. The example concerns the *Houaiss* entry for the word *sapato* 'shoe', given in Figure 2. After remarking on the obscure etymology of the word, the entry lists a number of cognates in related languages[10]. The LMF encoding for *sapato* is given in Figure 10. In this case, the *Lexical Entry* is linked with an individual of the class *Cognate-Set* to which various different cognates are associated.

In the appendix we include a brief description of an entry that combines information from two different resources.

5. Encoding the *Houaiss* Examples in TEI

In this section, we will look at the TEI encoding based on the serialisation introduced in LMF Part 4, of some of the etymologies which we presented above. As outlined in section 2.3., the TEI guidelines provide an optimal combination of semantically rich constructs for LMF Part 3 etymological descriptions along with additional mechanism to mark-up the source text in

fusórios que se encontra no ventre das rãs' ("the dating given is for the sense, not recorded here, of 'type of infusoria found in the abdomen of frogs'")

[9]Note that within the LMF model there is provision for specifying that the link is between a particular sense of the French word and a particular Portuguese entry (although this is not shown here).

[10]*etim orig. obsc., assim como o é em esp. zapato (1140) 'calçado', o cat. sabata (1268) 'calçado', provç sabata, fr. savate (c1200 chavate) 'calçado velho', it. ciabatta (a1140) 'pantufa, chinelo'* ("origin obscure, as in Spanish *zapato* (1140) 'footwear', or Catalan *sabata* (1268) 'footwear', Provencal *sabata*, French *savate* (c1200 chavate) 'old shoes', Italian *ciabatta* (before 1140) 'slipper, (another kind of) slipper'")

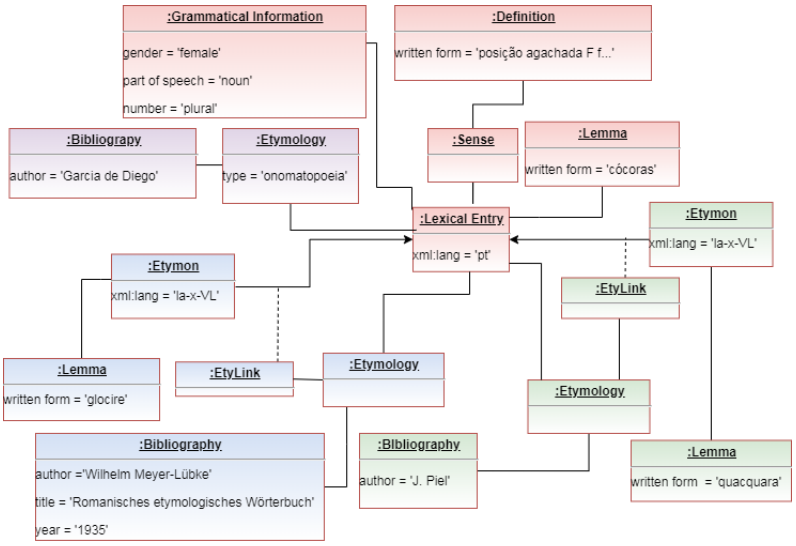


Figure 7: The LMF encoding for *cócoras*.

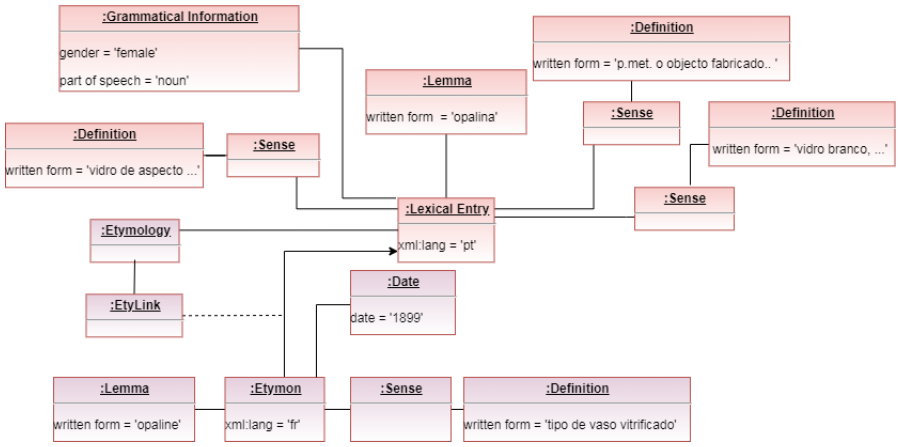


Figure 8: The LMF encoding for *opalina*.

sapato *s.m.* (sXIII cf. AGC) VEST calçado, ger. de sola dura, que cobre o pé, parcial ou completamente ✣ **s. abotinado** VEST *B* o que vai até à altura do tornozelo • **s. anabela** VEST sapato feminino de plataforma alta, ger. revestida de cortiça • **s. cara de gato** VEST *B* sapato de trança, feito em tecido de malha, e ger. com uma cara de gato desenhada a cores; charlote • **s. de defunto** *infrm.* promessa ou esperança demorada ou incerta • **s. de ferro** GIN aparelho de ferro fundido ou outro metal, que se ajusta ao pé para exercitar esp. as pernas e os músculos da região abdominal, nos exercícios de musculação • **s. de quarto** *P* chinelo • **s. raso** calçado que não cobre o peito do pé e que não tem tacão ou salto • **esperar s. de defunto** esperar por algo impossível ou de realização incerta • **saber onde aperta o s.** ter conhecimento da(s) causa(s) de um problema, de uma dificuldade ✣ ETIM orig. obsc., assim como o é em esp. *zapato* (1140) 'calçado', o cat. *sabata* (1268) 'calçado', provç *sabata*, fr. *savate* (c1200 *chavate*) 'calçado velho', it. *ciabatta* (a1140) 'pantufa, chinelo'; f.hist. sXIII *çapato*, sXV *sapato*

Figure 9: The Houaiss entry for *sapato*.

cases in which the encoding intends to be as faithful to the original representation of a lexical entry. The first example (See Listing 1[11]) shows the entry for *opalina*, which has been described in Section 4.. In the case of this simple etymological description, we have an <etym> element (with @type set to *borrowing*) containing a single <cit> for the one etymon to be described. This etymon, in turn, is associated with a language, a form, a bibliographical description, a cross-reference to another entry and a note[12].

[11]The full encoding of the example can be found under https://github.com/DARIAH-ERIC/lexicalresources/blob/master/Data/etymology/HouaissDictionary/opalina.xml

[12]Along with these components forming the core of the etymology of *opalina*, we can see how additional elements for marking-up labels and punctuation marks allow us to retain the linear flow of the original text.

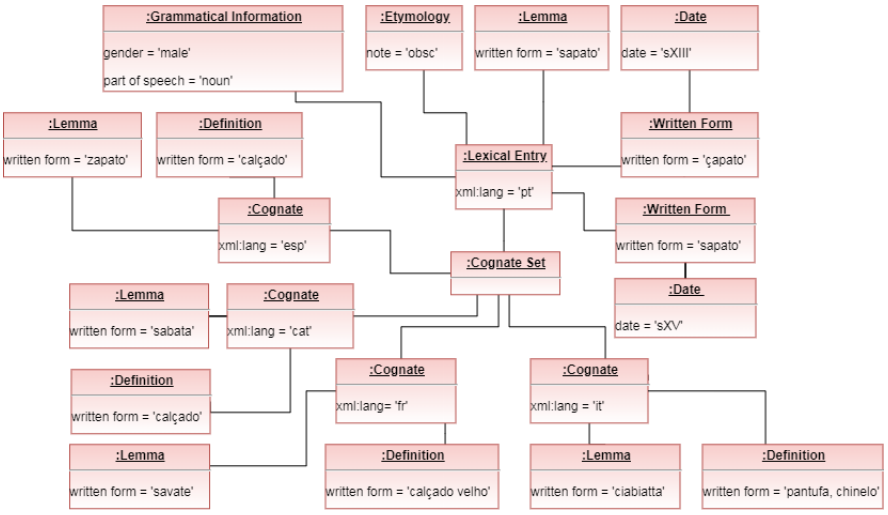


Figure 10: The LMF encoding for *sapato*.

Listing 1: TEI encoding of the etymology for the *opalina* entry.

```
<etym type="borrowing">
    <lbl>etim</lbl>
    <cit type="etymon">
        <lang expand="francês" norm="fr">fr.</lang>
        <form xml:lang="fr">
            <orth>opaline</orth>
        </form>
        <bibl type="attestation">
            <pc>(</pc><date>1899</date><pc>)</pc>
        </bibl>
        <pc>'</pc><gloss>tipo de vaso vitrificado</gloss><pc>'</pc
            >
        <pc>;</pc>
    </cit>
    <xr type="related">
        <lbl>ver</lbl>
        <ref type="entry">opal(i)−</ref>
    </xr>
    <note>a datação é para a ace., não registada aqui, género de infusó
        rios que se encontra no ventre das rãs</note>
    <pc>;</pc>
</etym>
```

The second example, given in Listing 2 corresponds to the encoding of the cognate set for *sapato*. We see here an immediate application of the recursive `<cit>` construct referring to the cognate set at the first level and to each cognate at the second level, with the appropriate @type attributes in place[13].

Listing 2: TEI encoding of the etymology for the *sapato* entry.

```
<etym>
    <lbl>etim</lbl>
    <note>etim orig. obsc., assim como o é em</note>
    <cit type="cognateSet">
        <cit type="cognate">
            <lang expand="espanhol" norm="es">esp.</lang>
            <form xml:lang="es">
                <orth>zapato</orth>
                <bibl type="attestation">
                    <pc>(</pc>
                    <date>1140</date>
                    <pc>)</pc>
                </bibl>
            </form>
            <pc>'</pc>
            <gloss xml:lang="pt">calçado</gloss>
            <pc>'</pc>
        </cit>
        <pc>,</pc>
        <cit type="cognate">
            <lang expand="catalão" norm="ca">o cat.</lang>
            <form><orth>sabata</orth></form>
            <bibl type="attestation">
                <pc>(</pc><date>1268</date><pc>)</pc>
            </bibl>
            <pc>'</pc>
            <gloss xml:lang="pt">calçado</gloss>
            <pc>'</pc>
        </cit>
        <cit type="cognate"> ... </cit>
        <cit type="cognate"> ... </cit>
        <pc>,</pc>
        <cit type="cognate"> ... </cit>
        <cit type="cognate"> ... </cit>
    </cit>
</etym>
```

6. Conclusions

Our main goal in this article has been to demonstrate the use and utility of these two new parts of LMF and to communicate the particular approach to etymological modelling which we have taken to the language resources community. This work presented in this paper is part of a broader experiment consisting of the analysis and conversion of different, heterogeneous, Portuguese lexical resources including legacy dictionaries such as the *DLPC* and *Houaiss* into computational formats such as TEI which is currently in progress. Our intention in future work is to compare

[13] Beyond the descriptive constructs we have seen in the previous example, we have here the provision of attestation seen as a bibliographical reference; the example also illustrates the use of the `<gloss>` element for the provision of translation equivalent in the working language. However, we haven't had space to discuss how these are represented in LMF.

different TEI based approaches, and in particular TEI Lex-0 (a streamlined version of the TEI standard for dictionaries) and LMF-in-TEI, to see which is more suited to different kinds of lexicographic resources. On the LMF side, we intend to continue building on the core model established for Part 3 and described in this article by taking into consideration a large number of diverse and wide-ranging case studies while further elaborating, in tandem, on the TEI serialisation of the other parts of LMF.

7. Acknowledgements

Fahad Khan and Ana Salgado were supported by the EU H2020 programme under grant agreements 731015 (ELEXIS – European Lexical Infrastructure). This research was also financed by Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020.

8. Bibliographical References

Bowers, J. and Romary, L. (2017). Deep encoding of etymological information in TEI. *Journal of the Text Encoding Initiative*, (10), August.

Budin, G., Majewski, S., and Mörth, K. (2012). Creating lexical resources in tei p5. a schema for multipurpose digital dictionaries. *Journal of the Text Encoding Initiative*, (3).

DLPC. (2001). *Dicionário da Língua Portuguesa Contemporânea, João Malaca Casteleiro (coord.), 2 vols.* New digital edition under revision, Ana Salgado (coord.).

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., and Soria, C. (2006). Lexical markup framework (lmf). In *International Conference on Language Resources and Evaluation-LREC 2006*.

Francopoulo, G. (2013). *LMF lexical markup framework*. Wiley Online Library.

Houaiss, A. (2001). *Dicionário Houaiss da Língua Portuguesa [CD-ROM]*. Rio de Janeiro, Ed. Objetiva.

Houaiss, A. (2015). *Grande Dicionário Houaiss da Língua Portuguesa*. Instituto António Houaiss Bloco Gráfico, Lda. Lisboa:Círculo de Leitores.

Khan, A. F. (2018). Towards the representation of etymological data on the semantic web. *Information*, 9(12):304, Nov.

Romary, L. and Tasovac, T. (2018). TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources. In *TEI Conference and Members' Meeting*, Tokyo, Japan, September.

Romary, L., Khemakhem, M., George, M., Bowers, J., Khan, F., Pet, M., Lewis, S., Calzolari, N., and Banski, P. (2019). Lmf reloaded. In *Asialex 2019*.

Romary, L. (2010). Using the TEI framework as a possible serialization for LMF. Rendering endangered languages lexicons interoperable through standards harmonization., August.

Romary, L. (2015). Tei and lmf crosswalks. *JLCL*, 30:47–70.

opalina [ɔpɐlínɐ]. *s. f.* (De *opala* + suf. *-ina*). **1.** Vidro translúcido, de aspecto semelhante ao da opala, que se usa no fabrico de objectos decorativos. *Jarra em opalina.* **2.** Objecto fabricado com essa espécie de vidro. *Colecção de opalinas.* **3.** Vidro espesso, de cor branca leitosa, utilizado no revestimento de paredes, tectos... **4.** *Zool.* Género (*Opalina*, Purk.) de protozoários ciliados, que compreende várias espécies.

Figure 11: The *DLPC* entry for *opalina*.

Salmon-Alt, S. (2006). Data structures for etymology: towards an etymological lexical network. *Bulletin de linguistique appliquée et générale*, 31:1–12.

Appendix

In the main body of the paper we have looked at how LMF can be used to encode lexical information that derives from a single source, in this case the *Houaiss* dictionary. However LMF is also intended to facilitate the integration of etymological information originating from different sources, while rendering provenance information as explicit as possible. To demonstrate this we will look at an example LMF entry that integrates information from two different Portuguese dictionaries: the aforementioned *Houaiss* and the *Dicionário da Língua Portuguesa Contemporânea(DLPC, 2001)*; the latter work constituting the first complete dictionary published by the *Academia das Ciências de Lisboa*. We will look at the entry for *opalina* from the *DLPC* (Figure 11) and combine the etymological information contained in that entry with that contained in the entry for the same word in *Houaiss* (Fig 5). Etymologies in *DLPC* entries are indicated in parentheses after the part of speech information and describe the origin of the word and the elements that go into its formation, along with the meanings of individual etymons when these are not obvious from the definitions in the entry. In the case of *opalina*, the *DLPC* etymology gives a derivation for the word from the noun *opala* and the suffix *-ina* but does not go into any further detail as to its origins. An LMF entry encoding the etymological information contained in both resources can be seen in Figure 12. Here the green boxes represent the etymological information from *DLPC* and the purple boxes the information deriving from *Houaiss*[14].

[14]Note that in the case of derivation relations when the etymological link is a three (or more) place relation we use numbers on the different individuals which are connected together to represent their ordering

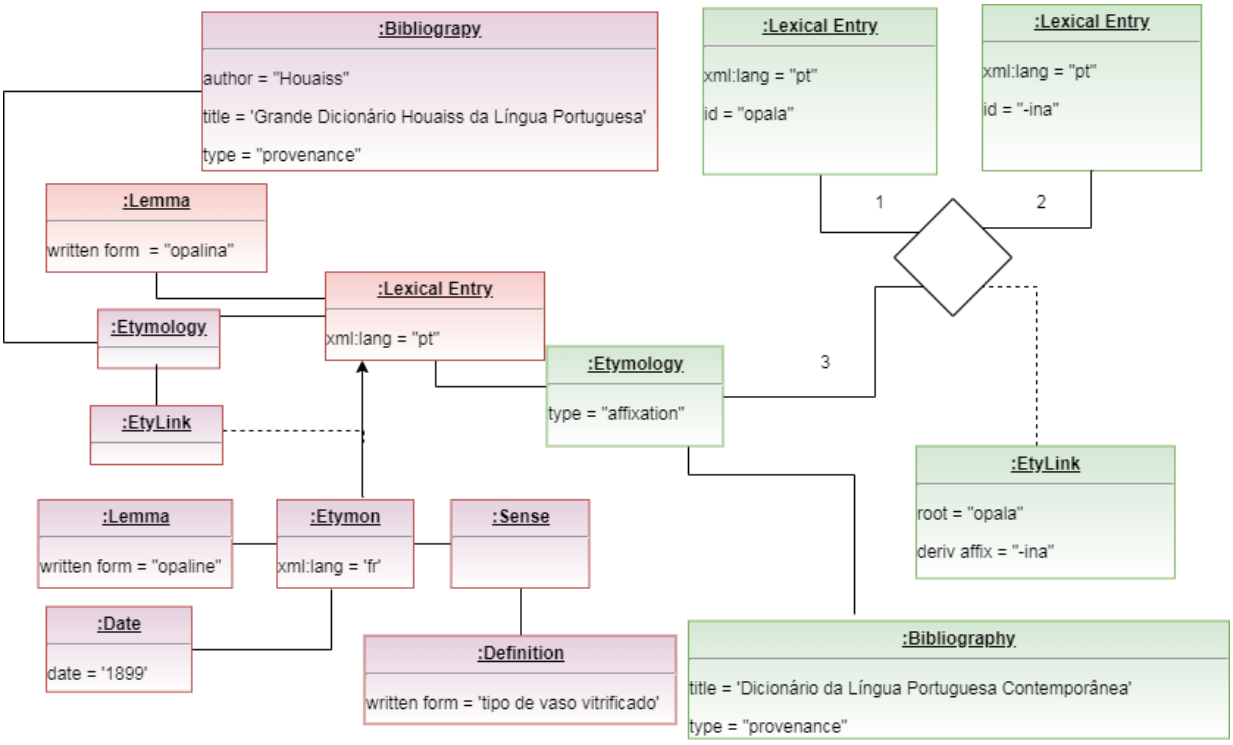


Figure 12: Combined entry for *opalina*.