TEI Lex-0 Etym – towards terse recommendations for the encoding of etymological information

Jack Bowers, Axel Herold, Laurent Romary, Toma Tasovac

Journal of the Text Encoding Initiative
Rolling Issue | 2022

TEI Lex-0 Etym: Toward Terse Recommendations for the Encoding of Etymological Information

Jack Bowers, Axel Herold, Toma Tasovac and Laurent Romary

Electronic version
URL: https://journals.openedition.org/jtei/4300
DOI: 10.4000/jtei.4300
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Jack Bowers, Axel Herold, Toma Tasovac and Laurent Romary, "TEI Lex-0 Etym: Toward Terse Recommendations for the Encoding of Etymological Information", *Journal of the Text Encoding Initiative* [Online], Rolling Issue, Online since 20 September 2022, connection on 14 January 2023. URL: http://journals.openedition.org/jtei/4300 ; DOI: https://doi.org/10.4000/jtei.4300

TEI Lex-0 Etym: Toward Terse Recommendations for the Encoding of Etymological Information

Jack Bowers, Axel Herold, Toma Tasovac, and Laurent Romary

ABSTRACT

The present paper describes the specific contribution of the TEI Lex-0 initiative, which aims to define a terser subset of the TEI Guidelines for the representation of etymological features in dictionary entries. Going beyond the basic provision of etymological mechanisms in the TEI Guidelines, TEI Lex-0 Etym proposes a systematic representation of etymological and cognate descriptions by means of embedded constructs based on the `<etym>` (for etymologies) and `<cit>` (for etymons and cognates) elements. In particular, given that all the potential contents of etymons are highly analogous to those of dictionary entries in general, the contents presented herein heavily reuse many of the corresponding features and constraints introduced in other components

of TEI Lex-0 for the encoding of etymologies and etymons. The TEI Lex-0 Etym model is also closely aligned with ISO 24613–3 on modeling etymological data and the corresponding TEI serialization available in ISO 24613–4.

INDEX

Keywords: Dictionary, etymology, TEI

ACKNOWLEDGEMENTS

The work described in this paper has benefited from a lot of interactions with all members of the DARIAH Working Group on Lexical Resources[1] and various colleagues from the EU project Elexis.

1. Introduction

1 Since the beginning of the Text Encoding Initiative and the first publication of the TEI Guidelines, the dictionary chapter has constantly been a reference model for representing structured lexical information. Still, as already observed (Romary and Wegstein 2012; Romary 2015), the TEI dictionary chapter lacks precision in offering univocal constructs for a given lexical phenomenon or feature. The case of etymological descriptions is unique in that it has remained underspecified. For instance, in the examples given in the TEI Guidelines (TEI Consortium 2022, ch. 9: Dictionaries),[2] etymological content is either left untagged or at best lightly annotated as mixed content; thus, more elaboration on this topic is needed. To address this need, the provisions described in this paper are being presented as the TEI Lex-0 Etym extension of the larger TEI Lex-0 initiative.

2 Building on recent efforts addressing etymology in TEI (Bowers and Romary 2017; Salmon-Alt 2006) in combination with the work carried out by the ISO (ISO 2021), TEI Lex-0 Etym defines a more restrained set of options for encoding any given single etymological phenomenon. The recommendations herein are designed to be able to handle born-digital as well as retro-digitized print sources, for which more flexible representation mechanisms may be needed. The present proposal is named after TEI Lex-0 (Romary and Tasovac 2018), an initiative launched in 2016 under the auspices of the DARIAH Working Group on Lexical Resources, which aims to define a pivot format for the integration and querying of heterogeneous TEI-based lexical resources.[3]

3 The scope of our proposal covers the usage of the following concepts central to etymological description:

- Structuring etymologies through ordering and (optionally) recursivity
- Typology of etymological processes
- Etymons, their forms, senses, and additional characterizing information
- Related forms (cognates, derivatives, and others)
- Temporality of etymological processes
- Bibliographical references in etymologies
- Prose description of etymological process and content
- Provenance, opinion, and conflicting/divergent etymological accounts

4 As we will demonstrate, inherent in all etymological descriptions are certain lexicographic objects, features, and relations. These relations are temporal in the case of etymons (representing a stage in the linguistic development of a lexeme), and/or spatial in the case of cognates (forms in a related language derived from the same etymological origin as given form). In order to concretely improve this long-neglected component of the TEI, and to facilitate the encoding and exploitation (search and retrieval) of etymological lexicographic data, herein we present a systematic representation of all descriptive components associated with these lexicographic objects, features, and relations.

5 Additionally, where the content is overlapping, the recommendations in this paper are intended to reuse those which are defined in other sections of TEI Lex-0. Any exceptions to this are due to specific needs of a given feature.

2. I. Overview of TEI Elements Considered in This Paper

6 In this section, we present the basic building blocks of creating structured etymological contents in a TEI dictionary. The components discussed are all pre-existing within the TEI schema; however, in order to best accommodate the concrete lexicographic needs common to both retro-digitized and born-digital etymological entries, some elements and attributes have been customized in the TEI Lex-0 schema.[4]

2.1 <etym>

7 The basic element within which all etymological content should be described is the etymology element <etym>. With a few modifications which are described below, the basis for the use of the <etym> element is the same as that described by Bowers and Romary (2017), which features three options for its placement in an entry:

- as a child of <entry>, when describing the history of the lexical entry as a whole
- as a child of <sense> for sense-based changes, or when such content is present in a legacy print source
- (in conjunction with one of the above) embedded (0..n) times within another <etym> to represent multiple ordered processes in sequence

8 Another key feature is that the attribute @type can be used on <etym> to specify etymological processes explicitly. If the process has subtypes, @subtype can also be used (see section Nested and Typed Structure). Finally, the certainty attribute @cert can be used in cases where the etymological description may not be certain. The default values of @cert are "high", "medium", "low", or "unknown"; editors could use any combination of these as needed.

2.2 Subcomponents of <etym>

9 Within <etym> the following elements can occur any number of times (some occur within each other):

- <cit>[5] with @type for complex descriptions of linguistic signs and their properties. The two most common usages in the context of etymological representation are etymons (cit[type="etymon"]) and cognates (cit[type="cognate"]).[6] We follow here the recent developments related to ISO standard 24613–3: see Khan and Bowers (2020). The element may carry the appropriate @xml:lang attribute to indicate the actual language of the etymon or cognate.
- <lang>, with the attributes @expand (full name of language for documentation purpose) and @norm (an encoded value according to BCP 47)[7] for marking up references to languages mentioned in the etymological process.

- `<date>`[8] for dating information (complementarity to `<lang>`).
- `<bibl>` and `<biblStruct>` for (complete) bibliographical references presented inline.
- `<ref type="bibl">` for pointers to bibliographical entries stored elsewhere, which may also contain a `@target` attribute when the bibliographic description is available.
- As an option depending on editorial practices, `<seg type="desc">` for spans of prose that do not represent any of the information types described above.
- `<note>` for editorial notes that are not part of the actual etymological description (see the previous discussion concerning`<seg type="desc">`).
- `<lbl>` to mark up short intertwining descriptive or connecting markers, particularly in cases of cross references (e.g., cf.andsee).
- `<xr>` (combined with `<lbl>` and `<ref>` as specified in TEI Lex-0: see Tasovac et al. 2018, ch. 7: Cross-references[9]) for cross-references to other lexical entries, forms, or senses, typically reflecting lexical-semantic relations.
- In cases where explicit etymological links have to be expressed, an additional `<link>` element may be used to link etymons or cognates as described in ISO 24613–3 and ISO 24613–4.

2.3 Basic Components of Etymons, Related Forms, and Other Components of Etymologies

10 Other than `<seg type="desc">`, `<bibl>`, `<date>`, and `<note>`, the rest of the most important components of an etymology, which are described in the following sections, are encoded as children of a typed `<cit>` element for describing etymons (`type="etymon"`) or cognates (`type="cognate"`) , both of which are discussed in detail in section 4.

11 The element `<cit>` can contain:

- `<form>` for describing the actual form corresponding to the intended etymon or cognate with related constraints (for example, must contain `<pron>` or `<orth>`) similar to those that apply in the general TEI Lex-0 specification. In specific cases, when the information is not provided at the etymon level, the element may carry the appropriate `@xml:lang` attribute to indicate the actual language of the form.

- `<gramGrp>` for providing the grammatical properties associated with the etymon or cognate (as in TEI Lex-0).
- `<lang>`[10] for mentioned names of the languages. When both a form and a language are provided, we recommend using `@norm` on `<lang>` with the same value as the language indication on `form/@xml:lang` or `cit/@xml:lang`.
- `<date>`[11] for dating information (period of occurrence of the etymon, whether attested or inferred).
- `<gloss>`[12] in the case of a simple equivalent or paraphrase in the working language of the dictionary.
- `<def>` for lexicographic definitions of the etymon or cognate.
- `<sense>` when the description of the etymon or cognate requires a structured semantic description (as in TEI Lex-0).
- `<usg>` for usage information (as in TEI Lex-0).
- `<xr>` (combined with `<lbl>` and `<ref>`) for additional lexical, etymological, or semantic relations, for example, "meronymOf."[13]
- `<ref type="bibl">` with a `@target` attribute for references to bibliographic entries described elsewhere in the encompassing document, and possibly `<bibl>` as an alternative, when no central bibliographical management is anticipated for the current dictionary.

12 As can be seen already, and in continuity with Bowers and Romary (2017), the TEI Lex-0 recommendation departs significantly from the traditional use of `<mentioned>` for representing etymons. There are two primary reasons for this. First, we wanted to provide the corresponding forms with the appropriate mechanisms to describe etymological forms (in the lexicographic sense) with a maximal degree of granularity and specificity: that is, it is essential to be able to qualify the nature of their orthographic or phonetic characteristics. Second, from a broader perspective, given that etymological entries may include a wide variety of lexical information (for example, glosses, grammar, and usage), we see etymological constructs (with `<cit>`) as simplified lexicographic entries[14] which need to closely mirror the structure of entries at large.

13 It should also be noted here that making a clear-cut decision as to what is to be considered a gloss or definition is not always straightforward, particularly in cases where the meaning is given by a paraphrase in the working language (for example, “hin- und herlaufen,” which is neither an equivalent lexeme nor a fully-fledged definition). Still, lexicographic practice has led us to keep the two possibilities and require encoders to document precisely the differentiation criteria they have used concerning these two elements. Moreover, in keeping with the general principles of TEI Lex-0, we strongly recommend that any complex semantic description associated with an etymon actually be embedded within a container `<sense>` element.

3. Structuring an Etymology

3.1 Minimal TEI Lex-0 Etymology Encoding (Flat, Non-typed)

14 The most fundamental requirement of any TEI encoding of etymological information should simply be to include this information inside the `<etym>` element.In marking up an etymological entry, there are several key structural decisions that will be up to the encoder and should conform to the source and target structure of the data itself. Minimal adherence to TEI Lex-0 Etym requires only that the data be encoded using the elements described above: all text content must be wrapped in the particular element(s) specified for their data type, with `<seg type="desc">` remaining an option depending on editorial practices. Optionally, users can include multiple layered `<etym>` elements which may also be typed. In the sections below we describe each basic possibility, their uses, and the specifics of their encoding.

15 In example 1, from Kluge’s etymological dictionary of German (Kluge 1975), we demonstrate the minimal encoding of the entry components.

16 “ Eingang m.

mhd. īnganc, nnl. ingang, dän. indgang, schwed. ingång: Lehnübersetzung des lat. introitus.

Aus dem ‘Hineingehen’ als Handlung ist die ‘Stelle, an der man ins Haus, in den Saal geht’ geworden, neuerdings auch die ‘Gesamtheit der eingegangenen Geschäftssachen, Mannschaften’ usw. Vgl. Zugang.” (Kluge 1975, 159)

Example 1. Minimal encoding of an etymological description. (Source: Kluge 1975).

```
<entry xml:id="Eingang" xml:lang="de">
```

```
<form type="lemma"><orth>Eingang</orth></form>
<gramGrp><gram type="gen">m.</gram></gramGrp>
<etym>
 <cit type="etymon" xml:lang="gmh">
  <lang expand="Mittelhochdeutsch" norm="gmh">mhd.</lang>
  <form><orth>înganc</orth></form>
 </cit>
 <cit type="cognate" xml:lang="nl">
  <lang expand="Neuniederländisch" norm="nl">nnl.</lang>
  <form><orth>ingang</orth></form>
 </cit>
 <cit type="cognate" xml:lang="da">
  <lang expand="Dänisch" norm="da">dän.</lang>
  <form><orth>indgang</orth></form>
 </cit>
 <cit type="cognate" xml:lang="sv">
  <lang expand="Schwedisch" norm="sv">schwed.</lang>
  <form><orth>ingång</orth></form>
 </cit>
 <lbl>Lehnübersetzung des</lbl>
 <cit type="etymon" xml:lang="la">
  <lang expand="Latein" norm="la">lat.</lang>
  <form><orth>introitus</orth></form>
 </cit>
 <note>Aus dem ‘Hineingehen’ als Handlung ist
  die ‘Stelle, an der man ins Haus, in den Saal geht’
  geworden, neuerdings auch die ‘Gesamtheit der
  eingegangenen Geschäftssachen, Mannschaften’
  usw. Vgl. <xr type="related"> <ref type="entry">Zugang</ref>.
   <ref type="bibl">(Kluge, 1975) p. 159</ref></xr></note>
</etym>
</entry>
```

17 Note that the use of etymons and cognates will be discussed in detail in section 4.

18 In example 1, all the main data components are tagged in the same relative location as in the printed source. Note also that even though no explicit typology is used here (that is, in the absence of `etym/@type`, this encoding still contains a significant amount of machine-retrievable

information pertaining to the etymological processes involved. For example, in <lbl> there are the words “Lehnübersetzung des” loan translation of (also known as “calque”) and “lat. introitus” of the Latin introitus which contains the source language. Additionally, the presence of the Middle High German language (<lang>mhd.</lang>) would enable researchers to infer the process of inheritance into Modern German.

19 In any given project where terminology is consistent,[15] and where the proper references to parent language stages are present, the presence of such information will enable a certain degree of machine-retrievability even without adding any additional structure in the TEI encoding.

3.2 Ordering of Embedded Etymologies to Encode Chronology

20 Where an entry has an etymology with multiple stages, the embedded <etym> elements should be ordered so that the element at the highest position in the hierarchy represents the most recent stage and the one at the lowest position represents the oldest stage.

Example 2. Embedded <etym> stages: source ordered (most to least recent).

```
<etym>Inherited from Middle English X
 <etym>from Old English Y
  <etym>which was borrowed from Latin Z
   <etym>which was from the Proto Italic Q
    <etym>from Proto Indo-European Ʊ</etym>
   </etym>
  </etym>
 </etym>
</etym>
```

21 While the structure of the source in example 2 is the ideal case in that the ordering of the contents is also from the most to least recent, there may be data sources or editorial preferences that present the etymology in a reverse, or indirect, order. In such cases, this XML hierarchy structure can nonetheless be maintained, while the information can be presented in an inverse or even nonlinear chronological order while still remaining structured and predictable so that a given XML <etym> layer will correspond programmatically to a relative stage in the chronology. An instance of such a case is shown in example 3:

Example 3. Embedded stages: source ordered (most to least recent).

```
<etym>
 <etym>
  <etym>
   <etym>ultimately from Proto Indo-European Ʊ</etym>
  which was from the Proto Italic Q
  </etym>
 borrowed from Latin Z
 </etym>
inherited from Middle English X
</etym>
```

3.3 Nested and Typed Structure

22 The nested <etym> structure was introduced by Bowers and Romary (2017) and allows for the recursion of an <etym> for the purposes of encoding multiple stages of an etymology and/or where an etymological change is complex and is inherently comprised of multiple interacting processes. Typing can of course be done using the @type attribute, and if a project's taxonomy/ontology of etymological processes has subtypes (i.e., if calques or loan translations are a subtype of the process borrowing), the @subtype attribute can also be used.

23 Re-examining example 1 above from Kluge (1975), we can see that it actually has two potential etymological layers which can be further structured. The entry implicitly states that: a) the word is inherited from Middle High German īnganc; and b) a common ancestor of both German Eingang as well as the cognate forms in other Germanic languages share the same source. The entry explicitly states that the word in each of those languages is a loan translation (Lehnübersetzung), that is to say a calque,from the Latinintroitus.In this alternate encoding of the etymology portion of this entry, the chronological ordering of the etymological processes is represented in the structure as described in the previous section, from most recent (inheritance from Middle High German) on the uppermost <etym> to least recent (loan translation/calque from Latin) on the lowermost (embedded) <etym>.

Example 4. Alternate encoding of the etymology of "Eingang" from example 1. (Source: Kluge 1975).

```
<etym type="inheritance">
 <cit type="etymon" xml:lang="gmh">
  <lang expand="Mittelhochdeutsch" norm="gmh">mhd.</lang>
  <form><orth>īnganc</orth></form>
```

```
</cit>
<!-- cognates here -->
<etym type="borrowing" subtype="calque">
 <lbl>Lehnübersetzung des</lbl>
 <cit type="etymon" xml:lang="la">
  <lang expand="Latein" norm="la">lat.</lang>
  <form><orth>introitus</orth></form>
 </cit>
</etym>
<note>Aus dem ‘Hineingehen’ als Handlung ist die
 ‘Stelle, an der man ins Haus, in den Saal geht’
 geworden, neuerdings auch die ‘Gesamtheit
 der eingegangenen Geschäftssachen, Mannschaften’
 usw. Vgl. <xr type="related"> <ref type="entry">Zugang</ref>.
  <ref type="bibl">(Kluge, 1975) p. 159</ref></xr></note>
</etym>
```

24 This nested structuring is not always possible, especially in cases of retro-digitized print sources where the order in which the forms and data are presented may limit or prohibit the use of nesting in any kind of systematic way. See Bowers and Romary (2017) for more in-depth discussion of embedding `<etym>`.

3.4 Descriptions and Prose

25 Prose descriptions of the etymological components and processes can be represented in several different ways according to the editorial approach. They could be left alone untagged (text can be placed directly in `<etym>`); specifically annotated with `<lbl>` or `<note>` elements when appropriate; or, when one wants to uniformly embed all linguistic descriptions at the same encoding level in the XML tree, with the systematic use of a `<seg type="desc">` element. In some cases the descriptions found in print dictionaries may occur in multiple discontinuous parts interrupted by examples or other structured content. In such cases, where the editors wish, the attribute `@part` can be used with a value of `"I"` initial, `"M"` medial, or `"F"` final,[16] where the value `"M"` may be used any number of times as needed. Example 5 shows this and some of the other primary elements we have listed above that are used in etymologies. To the left is the print source with the TEI encoding on the right. Note that, while dating information is indeed conceptually

part of a description of the etymology, when possible, it is ideal to keep <date> as a child of <etym> (though <date> is allowed within <seg> where it is too burdensome, or deemed unnecessary to separate it from the prose).

26 “Etymologie

Seit dem 18. Jh. belegt, auf fickfacken ‘hin- und herlaufen’ zurückgeführt: evtl. Auch auf fnhd. Fatzen ‘spotten, zum Narren halten’ zurückführbar (vgl. Pfeifer 2014:329)”Kluge 1975

Example 5. Use of <lbl>, <date>, <bibl>, and <seg type="desc"> with discontinuous prose. (Source: Kluge 1975).

```
<etym>
 <lbl>Etymologie</lbl>
 <date>Seit dem 18. Jh.</date>
 <seg type="desc" part="I">belegt, auf</seg>
 ...
 <seg type="desc" part="M">zurückgeführt: evtl. Auch auf</seg>
 ...
 <seg type="desc" part="F">zurückführbar</seg>
 <xr><pc>(</pc><lbl>vgl.</lbl><ref>Pfeifer 2014:329</ref><pc>)</pc></xr>
</etym>
```

4. Etymons and Other Forms

27 The basic component of an etymology is an etymon, which is often represented by a form and which may include other information typical of any lexical entry: for example, a language name, grammatical properties, usage descriptions, semantic descriptions, or bibliographic sources. Etymons are encoded in <cit type="etymon"> and are used analogously to the organization of <entry> both conceptually and structurally as shown by the side-by-side comparison of Entry structure and contents (example 6) and basic etymon (example 7).

Example 6. Entry structure and contents. (Source: Bowers 2020).

```
<entry xml:id="ntuchi" xml:lang="mix">
 <form type="lemma">
  <orth>ntuchi</orth>
  <pron notation="ipa">nduʧí</pron>
 </form>
```

```
 <gramGrp>
  <pos>noun</pos>
 </gramGrp>
...
</entry>
```

Example 7. Basic etymon. (Source: Bowers 2020).

```
<cit type="etymon" xml:lang="mix">
 <form>
  <orth corresp="#ntuchi">ntuchi</orth>
 </form>
 <gloss xml:lang="en">bean</gloss>
 <gloss xml:lang="es">frijol</gloss>
</cit>
```

4.1 Specific Types of Etymon Structures

28 The data structure of etymons can vary in certain ways according to the specifics of the conceptual content, purpose, and/or sources. A few examples of such variation are:

- if the etymon is based in external sources
- if it expresses a semantic change (but nota form change)
- if it expresses provenance while no form attested in the source language is provided

29 Below we demonstrate such scenarios and their encoding.

4.2 Linking Forms to External References

30 If an encoder wants to link a form to an existing external resource, it can be done using a `@corresp` attribute on the `<form>` element.

Example 8. Linking the form associated with an etymon to an external reference.

```
<cit type="etymon" xml:lang="und-x-pie">
 <form corresp="http://example.org/uekw.htm">
  <pron>u̯ekʷ-</pron>
 </form>
</cit>
```

4.3 Etymons in Semantic Changes and Polysemy

31 In certain cases (such as in an etymology describing a semantic change, resulting in polysemy), the etymon may consist of only a semantic description with no form. This is possible because in cases of polysemy, the form of the new meaning/lexical item remains the same as the headword of the entry. In the encoding in example 9, the `@corresp` attribute added to the `<cit type="etymon">` points to the `@xml:id` value of the source sense. Example 9 shows the etymon with only sense change; example 10 shows the entry to which the etymon is referring.

Example 9. Etymon with only sense change.

```
<entry xml:id="body-face">
 <form type="lemma">
  <orth xml:lang="mix">nuu</orth>
  <pron notation="ipa" xml:lang="mix">n`ŭű</pron>
 </form>
 …
 <sense xml:id="face-PRIME" n="1">
  <usg type="domain">Anatomical Structure</usg>
  …
  <cit type="translation">
   <form>
    <orth xml:lang="en">face</orth>
   </form>
  </cit>
  …
  <sense xml:id="face-regionOfLM" n="2">
   <usg type="domain">Space</usg>
   <usg type="domain">Spatial Relations</usg>
   <gramGrp>
    <pos>relationNoun</pos>
    <colloc norm="landmark">LM</colloc>
   </gramGrp>
   <def xml:lang="en">The front of (sth).</def>
   …
   <cit type="example">
    <quote xml:lang="mix">nuu ve'e</quote>
    <cit type="translation">
```

```
      <quote xml:lang="en">the front of the house</quote>
     </cit>
    </cit>
    …
    <!-- <etym> goes here -->
   </sense>
  …
  </sense>
 </entry>
```

Example 10. Entry to which etymon in example 9 is referring.

```
<cit type="etymon" corresp="#face-PRIME">
 <sense>
  <usg type="domain">Anatomical Structure</usg>
  <gloss xml:lang="en">face</gloss>
  <xr type="meronymy">
   <lbl xml:lang="en">as in:</lbl>
   <ref type="sense" target="#body-face" xml:lang="en">part of the body</ref>
  </xr>
 </sense>
</cit>
```

4.4 Etymons with Provenance Information

32 In other cases, an <etym> may contain only the name of a language, expressing provenance (in a loanword or inheritance), or possibly simply a date corresponding to the time or first attested usage. Note that, as discussed above, typing on <etym> is optional; however, it is shown in example 11 and example 12 to complement these particular etymons. The use of the @xml:lang attribute on the <cit> level defines the etymon as being from the given language, despite not having a form therein. The inclusion of @xml:lang on <lang> therein distinguishes the working language (which is for the reader and is not actually part of the etymon) from the language of the etymon.

Example 11. Borrowing. Etymon with only provenance, without forms. (Source: Kluge 1975).

```
<etym type="borrowing">
 <lbl>aus</lbl>
 <cit type="etymon" xml:lang="sl">
```

```
  <lang xml:lang="de" expand="Slowenisch" norm="sl">slow.</lang>
 </cit>
</etym>
```

Example 12. Inheritance. Etymon with only provenance, without forms. (Source: Kluge 1975).

```
<etym type="inheritance">
 <cit type="etymon" xml:lang="gmh">
  <lang xml:lang="de" expand="Mittelhochdeutsch" norm="gmh">mhd.</lang>
  <ref type="bibl">Lexer Wb. III 324</ref>
 </cit>
</etym>
```

4.5 Etymons with Simple Dating Information

33 The information about etymons can sometimes be so reduced as to contain no source form information. This is the case in example 13, in which the author of the etymological description just wanted to record the actual period or date of the first occurrence of the etymon with limited etymological background apart from the temporal information itself (or because the form might be obvious for the reader).

34 »ins engl. Nach Trench 24 im 16. Jh. Mit dem hauptbegriffe des geschmückten gekommen; der ursprung ist sehr zweifelhaft«

Example 13. Etymon with date but not form information.

```
<cit type="etymon">ins engl.
 <xr>
  <lbl>nach</lbl>
  <ref type="bibliography">Trench 24</ref>
 </xr>
 <date>im 16. Jh.</date>
 <def>mit dem hauptbegriffe des
 geschmückten gekommen</def><pc>;</pc>
 der ursprung ist sehr zweifelhaft
 <xr>; s.v. /brave/ in Mueller (1878)</xr></cit>
```

4.6 Variants of Etymological Forms

35 As mentioned, in encoding any type of forms in etymologies, `<form>` and its child elements behave the same way as when they occur on the level of the main entry. This structure is necessary for encoders of etymological dictionaries because it is common to find multiple variants and/or inflected forms of the same etymon that cannot be listed separately, as they correspond to the same definition and/or other key pieces of information. Thus variants of etymons or other forms (cognates; see next section) in an etymology should be represented in accordance with the recommendation of the TEI Lex-0 Forms section (Bański, Bowers, and Erjavec 2017, sec. 2). [bad link to item:] shows two such occurrences from print dictionaries.

36 “ Etymologie

mhd. vreten, vretten, vraten ‘entzünden; wundreiben; herumziehen; quälen; plagen’ (vgl. Lexer 1878 III: 502) ” Lexer 1878)

Example 14. Variants of the Middle High German etymon “vreten, vretten, vraten” (Source: Lexer 1878).

```
<etym>
 <cit type="etymon" xml:lang="gmh">
  <lang>mhd.</lang>
  <form type="variant">
   <orth>vreten</orth>
  </form>
  <form type="variant">
   <orth>vretten</orth>
  </form>
  <form type="variant">
   <orth>vraten</orth>
  </form>
  <gloss>entzünden</gloss>;
  <gloss>wundreiben</gloss>;
  <gloss>herumziehen</gloss>;
  <gloss>quälen</gloss>;
  <gloss>plagen</gloss>
 </cit>
 <bibl><lbl>vgl.</lbl>
  <title>LEXER</title>
  <date>1878</date>
```

```
    <edition>III</edition>
    <citedRange>502</citedRange></bibl>
  </etym>
```

4.7 Cognates

37 Cognates are forms asserted as being related in some way to the lexical entry and/or the etymon and are a ubiquitous feature of etymological dictionaries. Cognates are essentially lexical items in a language that share an etymological source. The structure of a basic representation of cognates mirrors that of etymons and uses the same `<cit>` structure, with the difference that the value of `@type` should be "cognate." Note that in example 15, `<ref type="bibl">` is used instead of `<bibl>` because in the project from which the examples are taken all bibliographical sources are listed in the header with `@xml:id`s.

Example 15. Collection of cognates taken from various external sources. (Source: Bowers 2020).

```
  <cit type="cognate" xml:lang="mig">
   <lang>Chalcatongo Mixtec</lang>
   <usg type="geographic">
    <placeName>San Miguel El Grande</placeName>
   </usg>
   <form><pron notation="trans-macaulay-mig">šinì</pron></form>
   <ref type="bibl" target="#Macaulay-ChalcatongoMixtec-1996">(Macaulay, 1996)</
ref>
  </cit>
  <cit type="cognate" xml:lang="miy">
   <lang>Ayutla Mixtec</lang>
   <form><pron notation="trans-hill-1990-miy">shīhih</pron></form>
   <ref type="bibl" target="#Hills-AyutlaMixtec-1990">(Hills, 1990)</ref>
  </cit>
  <cit type="cognate" xml:lang="miz">
   <lang>Coatzospan Mixtec</lang>
   <form><pron notation="trans-smll-miz">rkɨ</pron></form>
   <ref type="bibl" target="#Small-CoatzospanMix-1990">(Small, 1990)</ref>
  </cit>
  <cit type="cognate" xml:lang="smd">
   <lang>San Martín Duraznos</lang>
   <form><pron notation="ipa">ʃīŋī</pron></form>
```

```
 <ref type="bibl" target="#Padgett-2017">(Padget, 2017)</ref>
</cit>
```

4.8 Cognate Sets from a Common Bibliographic Source

38 In some cases there may be a list of cognates in a print or born-digital source that are all from a single bibliographic source, or are at least presented in a source etymology as a set or list. Often these would have some kind of referential function word or abbreviation: for example, "cf. …" In such cases it may be desirable to present the list of cognates as the source intended and thus group them in a single wrapper `<cit type="cognateSet">`.

39 In the case of a cognate set where there is a referential function word or abbreviation, it should be tagged with `<lbl>` and included as a child of `<cit type="cognateSet">`, preceding the etymons (see example 16). Where there is a common bibliographic source, `<bibl>` or `<ref type="bibl">` (if the bibliographic sources are already declared elsewhere) should be a child of the `<cit type="cognateSet">` and placed after the given forms (see example 16).

Example 16. Template for a set of cognates (`"cognateSet"`) from a single bibliographic source.

```
<cit type="cognateSet">
 <lbl>Cf.</lbl>

 <cit type="cognate" xml:lang="ffr">
  <form><orth>...</orth></form>
 ….
 </cit>
 <cit type="cognate" xml:lang="und-x-pom">
  <form><orth>...</orth></form>
 ….
 </cit>
 <cit type="cognate" xml:lang="und-x-opd">
  <form><orth>...</orth></form>
 ….
 </cit>
 ……
 <bibl>Bibl Source Here</bibl>
</cit>
```

4.9 Descendant and Derivative Forms

40 Certain etymological dictionaries may include descendant and/or derivative forms which were derived from the headword.

41 In an entry where the lemma is a form which has been derived from another lexical item, in the etymology section, the typology would be "derivation" and the source term from which it was derived would be the etymon. However, in cases such as the one shown in example 17, in which derivative forms are presented as related entries to the headword prior to the etymology section proper, these should be encoded as embedded `<entry>` elements according to the TEI Lex-0 Guidelines (Tasovac et al. 2018, ch. 3: Entries),[17] and the value of the `@type` attribute should be "derivative."

42 " amārus 'bitter' [adj. o/ā] (Pl.+)

Derivatives: amārilūdō 'bitterness' (Varro+), amāror[m.] 'bitter taste' (Lucr.+).

PIt. *o/am-?

PIE *h2h3m-ro-? IE cognates: Skt. amlá- 'sour, acid,' OIc. apr 'sharp, cold,' OE ampre 'sour one,' MDu, amper 'bitter, sour' < PGm. *am(p)ra- 'sour'; ? OIr. om 'raw,' W. of possibly <*h2h3 -emo-,Skt. āmá- [adj.] 'raw, uncooked,' Gr. ωμός 'raw,' Arm. howm <*h2eh3mo-* "de Vaan 2008

Example 17. Derivatives. (Source: de Vaan 2008) .

```
<entry>
 ...
 <lbl>Derivatives</lbl><pc>:</pc>

 <entry type="derivative" xml:lang="la">
  <form><orth>amārilūdō</orth></form>
  <sense><gloss>'bitterness'</gloss></sense>
  <ref type="bibliography">(Varro+)</ref>
 </entry><pc>,</pc>

 <entry type="derivative" xml:lang="la">
  <form><orth>amāror</orth></form>
  <pc>[</pc><gramGrp><gen>m.</gen></gramGrp><pc>]</pc>
  <sense><gloss>'bitter taste'</gloss></sense>
  <ref type="bibliography">(Lucr.+)</ref>
 </entry><pc>.</pc>
```

```
</entry>
```

4.10 Cross-referenced Forms

43 Often within etymological discussions (as in other portions of a dictionary entry) there are references to forms which are etymons, cognates, or derivatives, but which, in a given specific context, are not actually being posited as such. Where an editor wants to specify the particular etymological relationship to the lemma or other form, that information can be specified using the attributes on the cross-reference `<xr>` and embedded reference `<ref>` elements.

44 This section overlaps with the TEI Lex-0 section on structured lexical references (Tasovac et al. 2018, ch. 7: Cross-references)[18] but its application in the context of etymological content requires us to clearly identify the contexts of usage, as well as the conceptual distinctions between the cross-referenced forms and the primary features of etymology markup discussed above.

45 In an etymology, the contexts in which a cross-reference form should be used include

- where a reference is made to another lemma or form from a separate entry
- where a reference is made to the lemma (or another form) in the synchronic entry
- where a reference is made to an etymon which is not the etymon for the actual lemma
- where a reference is made to a sense (corresponding to either of the above)

46 In the examples below we show specific use cases of cross-referenced forms in the context of etymological dictionary entries.

47 In this first case, the entry is for the Latin arcessō, -ere / accersō, -ere; within the etymology section of that entry there are references to the two lemma variants. The cross-references are encoded in `<xr type="crossReference" xml:lang="la"><ref type="entry">`. This format would also apply to external cross-references. In either case, the editors would also have the option of including a pointer to the given internal or external form(s) with the `@target` attribute on `<ref>`.

48 The stem occurs in two variants, accers- and arcess-, which suggests that...

Example 18. Cross-reference to lemma in etymology section.

```
...<xr type="related" xml:lang="la"><ref type="entry">accers-</ref></xr> ...
<xr type="related" xml:lang="la"><ref type="entry">arcess-</ref></xr> ….
```

49 The following example (from the same entry as above) contains a reference to an etymon of the lemma/entry itself, but this is a supplementary instantiation of the given etymon which occurs in the context of a discussion of a particular phenomenon (for example, a phonetic change[19]). In such examples as the following, the use of `<xr>` encodes an important conceptual distinction in the data as it allows `<cit type="etymon">` to be reserved for the form(s) in the given entry.

50 ....Nussbaum 2007b gives two more arguments for regarding accersoas original: the noun dorsum → dossum shows a phonetic change of...

Example 19. Cross-referenced forms for etymons not pertaining to the lemma.

```
  ...the noun
  <xr type="related" subtype="etymon" xml:id="etym-dorsum" xml:lang="la"><ref
type="entry">dorsum</ref></xr>
  <lbl>→</lbl>
  <xr type="related" prev="#etym-dorsum" xml:lang="la"><ref type="entry">dossum</
ref></xr>
shows a phonetic change of...
```

51 Finally, we have an example in which there is a cross-reference to a sense of an external entry (example 20). This is also encoded as `<xr type="crossReference">` but differs in that it is a reference to the sense of a given entry; thus we use `<ref @type="sense">` and the embedded `<gloss>` within, which needs to have the `@xml:lang` to distinguish from the value declared at the `<xr>` level. In this case there is also a form (included here in a separate `<ref type="sense">`); however, it is also possible that a cross-reference to a sense could occur without an accompanying form.

52 ... a verb in -cesso meaning “go get”would be favoured by its semantic neighbours...

Example 20. Cross-reference to a sense from an external entry.

```
  ….a verb in
  <xr type="related" xml:lang="la"><ref type="sense">-cessō</ref>
   <lbl>meaning</lbl>
   <pc>'</pc><gloss xml:lang="en">go get</gloss><pc>'</pc></xr>
would be favoured by its semantic neighbours...
```

5. Conclusion

53 In this paper, we have provided a comprehensive view of the core proposals from the TEI Lex-0 Etym initiative with the expectation that its powerful and structured representations will serve both to provide guidance for the encoding of etymological content in TEI, for which little precedent is available, and to facilitate the interoperable encoding of a vast variety of potential etymological features found both in print (for retro-digitization) and born-digital contents. Still, as we know from the TEI Guidelines, this must remain a work in progress, since future users of the TEI Lex-0 Guidelines are likely to come up with new issues and change proposals. The online management of the whole TEI Lex-0 initiative as an open-source project makes it possible for anyone to contribute and help improve the specification and documentation, which are already in use in the EU project Elexis as the default pivot format for lexical data integration.

54 Finally, additional work may indeed be necessary to achieve more precise and stable ontologies for typing etymological processes, as well as for qualifying cross-references in etymological contexts.

6. Appendix: Fully Encoded Examples

55 In the following examples, we have tried to illustrate interesting cases of etymological processes that show how TEI Lex-0 Etym can seamlessly take into account a variety of situations. All examples have been validated and included in the TEI Lex-0 GitHub environment.

6.1 Embedded Senses, Metaphor, and Compounding

56 Example 21 shows a case of an embedded sense from the Mixtepec-Mixtec TEI dictionary (Bowers 2020) in which the lemma form xini ve'e is a compound with one component that is metaphorical in nature. The portion of the etymology that is metaphorical (`<etym type="metaphor">`) is embedded within the `<etym type="compounding">`, and as it is relevant to the process of metaphor, within the `<etym type="metaphor">` is the domain (`<usg type="domain">`).

Example 21. Embedded senses, metaphor, and compounding. (Source: Bowers 2020).

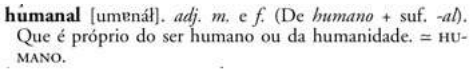
```
<sense>
 <usg type="domain">Architecture</usg>
 <cit type="translationEquivalent" xml:lang="en">
  <form>
```

```
    <orth>ceiling</orth>
   </form>
  </cit>
  <cit type="translationEquivalent" xml:lang="es">
   <form>
    <orth>techo</orth>
   </form>
  </cit>
  <etym type="compounding">
   <etym type="metaphor">
    <cit type="etymon" xml:lang="mix">
     <form type="lemma" corresp="#body-head">
      <orth>xiní</orth>
     </form>
     <gloss xml:lang="en">head</gloss>
     <gloss xml:lang="es">cabeza</gloss>
     <usg type="domain">Anatomy</usg>
    </cit>
   </etym>
   <cit type="etymon" xml:lang="mix">
    <form type="lemma" corresp="#house">
     <orth>ve'e</orth>
    </form>
    <gloss xml:lang="en">house</gloss>
    <gloss xml:lang="es">casa</gloss>
   </cit>
  </etym>
 </sense>
```

6.2 Derivational/Morphological Analysis

57 In example 22, for the Portuguese entry húmanal,we have a case of derivation (labeled with a `@type` attribute) in which the suffix -alis attached to the noun humanoto create the attributive adjective. Given that derivation can occur in a wide variety of morphological ways, the specific type of derivation is labeled with a `@subtype` attribute.

Figure 1. .


húmanal [umɐnál]. *adj. m.* e *f.* (De *humano* + suf. *-al*). Que é próprio do ser humano ou da humanidade. = HUMANO.

Example 22. Derivational/morphological analysis. (Source: de Vaan 2008).

```
<entry xml:lang="pt">
 <form type="lemma">
  <orth>húmanal</orth>
  <pron notation="ipa">umɐnáł</pron>
 </form>
 <gramGrp>
  <gram type="pos">adj.</gram>
  <gram type="gen">m.</gram>
  <lbl>e</lbl>
  <gram type="gen">f.</gram>
 </gramGrp>
 <etym type="derivation" subtype="suffixalDerivation">
  <pc>(</pc>
  <seg type="desc">De</seg>
  <cit type="etymon" xml:lang="pt">
   <form>
    <orth>humano</orth>
   </form>
  </cit>
  <lbl>+</lbl>
  <cit type="etymon" xml:lang="pt">
   <gramGrp>
    <gram type="pos" expand="suffix">suf.</gram>
   </gramGrp>
   <form>
    <orth extent="suff">-al</orth>
   </form>
  </cit>
  <pc>)</pc>
 </etym>
 <!-- rest of entry here -->
</entry>
```

6.3 Phonological Changes

58 It is very common in etymological dictionaries to have discussions about sequences of sound changes. These most often take place in the context of running prose. As described above, prose can be represented using `<seg type="desc">`, and where interrupted with etymons (or other content), the `@part` attribute can be used. The phonetic or phonological units described as having undergone particular changes are represented the same way as full word forms, using `<cit type="etymon">`. In order to attribute a unit's particular place in the sequence of sound changes, the `@prev` and `@next` attributes can be used to point to the `@xml:id` of the previous or next form in the diachrony. Example 23 illustrates these mechanisms.

59 Others have proposed an etymology *ad-arti- with intervocalic * d becoming l ; the spelling allerswould then be analogical to sollers

Example 23. Phonological changes. (Source: de Vaan 2008).

```
  <etym>
   <seg type="desc" part="I">Others have proposed an etymology</seg>
   <cit type="etymon" xml:id="ad-arti-" xml:lang="und-x-pie"><form><orth>*ad-
arti-</orth></form></cit>
   <etym corresp="#ad-arti-">
    <seg type="desc" part="I">with intervocalic</seg>
    <cit type="etymon" xml:id="c1" next="#c2"><form><orth>d</orth></form></cit>
    <seg type="desc" part="F">becoming</seg>
    <cit type="etymon" xml:id="c2" prev="#c1"><form><orth>l</orth></form></cit>
    <pc>;</pc>
   </etym>
   <seg type="desc" part="F">the spelling
    <xr type="crossReference" xml:lang="la"><ref type="entry">allers</ref></xr>
    would then be
    <xr type="crossReference" xml:lang="la"><lbl>analogical to</lbl>
     <ref type="entry">sollers</ref></xr><pc>.</pc></seg>
  </etym>
```

60 Note also that this entry contains an embedded etymology, and it is distinguished in the data structure according to the portion that belongs directly to the author, and the portion which he is ascribing to "others."

6.4 Multiple and/or Conflicting Etymological Accounts

61 In many sources there can be multiple, sometimes conflicting, accounts for an etymology. In these cases nested etymologies should be used, the top layer being reserved for the editorial descriptions, and any number of separate <etym>s can be included therein.

62 “ According to Untermann 2000, Latin *all-was probably borrowed from Sabellic, since Latin does not have this word in its lexicon. For a word only occurring in glosses, this is of course possible. Others have proposed an etymology *ad-arti-with intervocalic d becoming l; the spelling allerswould then be analogical to sollers. ”de Vaan 2008

Example 24. Multiple and/or conflicting etymological accounts. (Source: de Vaan 2008).

```
  <etym>
   <!-- PIt, PIE etymons-->
   <!-- cognates -->
   <etym cert="medium">
    <!-- Lat. sollers < *soti-arti- to sollus 'entire'; al(l)ers < *all-arti- to
O. alio- 'entire'. -->
    <seg type="desc" part="I">According to
     <ref type="bibliography">Untermann 2000</ref>,
    </seg>
    <xr type="crossReference"><lang>Latin</lang>
     <ref xml:lang="la">*all-</ref></xr>
    <seg type="desc" part="M">was probably
    borrowed from</seg>
    <cit type="etymon" xml:lang="und-x-sabe1249"><lang norm="und-x-
sabe1249">Sabellic</lang></cit>,
    <seg type="desc" part="F">since
     <lang norm="la">Latin</lang>
    does not have this word in its
    lexicon. For a word only
    occurring in glosses, this is
    of course possible.</seg>
   </etym>

   <etym cert="medium"><seg type="desc" part="I">Others have proposed an
etymology</seg>
```

```
    <cit type="etymon" xml:id="ad-arti-" xml:lang="und-x-pie"><form><orth>*ad-
arti-</orth></form></cit>

    <etym corresp="#ad-arti-">
     <seg type="desc" part="I">with
     intervocalic </seg><cit type="etymon" xml:id="c1" next="#c2"><form><orth>d</
orth></form></cit>

     <seg type="desc" part="F">becoming</seg>
     <cit type="etymon" xml:id="c2" prev="#c1"><form><orth>l</orth></form></cit>
     <pc>;</pc>
    </etym>

    <seg type="desc" part="F">the spelling
     <xr type="related"><ref xml:lang="la">allers</ref></xr>
     would then be <xr type="crossReference"><lbl>analogical
      to</lbl> <ref xml:lang="la">sollers</ref></xr><pc>.</pc></seg></etym>
  </etym>
```

BIBLIOGRAPHY

Bański, Piotr, Jack Bowers, and Tomaž Erjavec. 2017. "TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms." In *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference*, edited by Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, and Vít Baisa, 485–94. Brno: Lexical Computing. https://hal.inria.fr/hal-01757108; https://elex.link/elex2017/proceedings-download/.

Bowers, Jack. 2020. "Mixtepec-Mixtec Digital Dictionary / Dicionario del Mixteco de Mixtepec." Last updated December 6, 2020. https://github.com/iljackb/Mixtepec_Mixtec/blob/master/MIX-Lexicon-TEI-Dict.xml.

Bowers, Jack, and Laurent Romary. 2017. "Deep Encoding of Etymological Information in TEI." *Journal of the Text Encoding Initiative* 10. https://journals.openedition.org/jtei/1643; doi:10.4000/jtei.1643.

Crist, Sean. 2005. "Toward a Formal Markup Standard for Etymological Data." *Paper presented at the LSA Annual Meeting, Oakland, CA, January 6–9*. Available at http://www.sean-crist.com/professional/publications/crist_etym_markup.pdf.

ISO (International Organization for Standardization). 2021a. "Language Resource Management — Lexical Markup Framework (LMF) — Part 3: Etymological Extension." ISO 24613–3:2021. Geneva: ISO. https://www.iso.org/standard/75410.html.

———. 2021b. "Language Resource Management — Lexical Markup Framework (LMF) — Part 4: TEI Serialization." ISO 24613–4:2021. Geneva: ISO. https://www.iso.org/standard/75411.html.

Khan, Fahad, and Jack Bowers. 2020. "Towards a Lexical Standard for the Representation of Etymological Data." In *Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD). La Svolta Inevitabile: sfide e prospettive per l'Informatica Umanistica [Proceedings of the 9th Annual Conference of the AIUCD...]*, edited by Cristina Marras, Marco Passarotti, Greta Franzini, and Eleonora Litta, 125–29. https://aiucd2020.unicatt.it/aiucd-Khan_Bowers.pdf; doi:10.6092/unibo/amsacta/6316.

Kluge, Friedrich. 1975. *Etymologisches Wörterbuch der deutschen Sprache*. 21. unveränderte Aufl. Berlin: De Gruyter.

Lexer, Matthias von. *Mittelhochdeutsches Handwörterbuch*. 3: VF - Z, Nachträge . (1876 - 1878).

Romary, Laurent. 2015. "TEI and LMF Crosswalks." *Journal for Language Technology and Computational Linguistics* (JLCL) 30 (1): 47–70. https://jlcl.org/content/2-allissues/6-Heft1-2015/3Romary.pdf. Author's version available at https://hal.inria.fr/hal-00762664. Version published in journal, 2015: https://jlcl.org/content/2-allissues/6-Heft1-2015/3Romary.pdf.

Romary, Laurent, and Toma Tasovac. 2018. "TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources." Abstract in "The 18th Annual TEI Conference and Members' Meeting: Book of Abstracts," edited by Tensho Miyazaki, 274–75. Tokyo: Center for Evolving Humanities, Graduate School of Humanities and Sociology, The University of Tokyo; International Institute for Digital Humanities. https://tei2018.dhii.asia/AbstractsBook_TEI_0907.pdf. Author's version of presentation available at https://hal.inria.fr/hal-02265312/.

Romary, Laurent, and Werner Wegstein. 2012. "Consistent Modeling of Heterogeneous Lexical Structures." *Journal of the Text Encoding Initiative* 3. http://journals.openedition.org/jtei/540; doi:10.4000/jtei.540.

Salmon-Alt, Susanne. 2006. "Data Structures for Etymology: Towards an Etymological Lexical Network." *Bulletin de Linguistique Appliquée et Générale* 31: 101–12. Author's version available at https://hal.archives-ouvertes.fr/hal-00110971/.

TEI Consortium. 2022. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.4.0*. Last updated April 19, 2022. https://tei-c.org/Vault/P5/4.4.0/doc/tei-p5-doc/en/html/.

Tasovac, Toma, Laurent Romary, Piotr Banski, Jack Bowers, Jesse de Does, Katrien Depuydt, Tomaž Erjavec, et al. 2018. *TEI Lex-0: A Baseline Encoding for Lexicographic Data. Version 0.9.1*, last updated March 24, 2021. DARIAH Working Group on Lexical Resources. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

de Vaan, Michiel. 2008. *Etymological Dictionary of Latin and the Other Italic Languages*. Leiden Indo-European Etymological Dictionary Series 7. Leiden: Brill.

NOTES

1 Accessed June 17, 2022, https://www.dariah.eu/activities/working-groups/lexical-resources/.

2 https://tei-c.org/Vault/P5/4.4.0/doc/tei-p5-doc/en/html/DI.html.

3 See the project's GitHub repository, accessed June 17, 2022, https://github.com/DARIAH-ERIC/lexicalresources.

4 TEI Lex-0 schema v. 0.9.1, last updated March 25, 2022, https://github.com/DARIAH-ERIC/lexicalresources/tree/master/Schemas/TEILex0.

5 For a more general discussion on the `<cit>` element's semantics and usage scope, see Romary and Wegstein (2012).

6 For a complete list of customized `@type` values on `<cit>` in TEI Lex-0, see Tasovac et al. 2018, sec. 12.1.19: `<cit>`, https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#TEI.cit.

7 IANA (Internet Assigned Numbers Authority), Language Subtag Registry, last updated March 2, 2022, https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry. BCP 47, https://tools.ietf.org/rfc/bcp/bcp47.txt .

8 The elements `<date>` and `<bibl>` can also occur within `<cit>`.

9 https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#cross-references.

10 Whereas `<lang>` is currently not allowed in `<cit>` in the TEI Guidelines (TEI Consortium 2022, Appendix C: Elements, `<cit>`, https://tei-c.org/Vault/P5/4.4.0/doc/tei-p5-doc/en/html/ref-cit.html), TEI Lex-0 introduced the change in its specification. We suggest that the TEI Guidelines integrate the change by making `<lang>` a member of model.entryPart.

11 Note that while `<date>` is not permitted as a direct child element of `<cit>` in the general TEI Guidelines (TEI Consortium 2022, Appendix C: Elements, `<cit>`, https://tei-c.org/Vault/P5/4.4.0/doc/tei-p5-doc/en/html/ref-cit.html), it is available in the TEI Lex-0 customization; for further details, see Tasovac et al. 2018, sec. 12.1.19: `<cit>`, https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#TEI.cit.

12 Note that while `<gloss>` is not permitted as a direct child element of `<cit>` in the general TEI Guidelines (TEI Consortium 2022, Appendix C: Elements, `<cit>`, https://tei-c.org/Vault/P5/4.4.0/doc/tei-p5-doc/en/html/ref-cit.html), it is available in the TEI Lex-0 customization; for further details, see Tasovac et al. 2018, sec. 12.1.19: `<cit>`, https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#TEI.cit.

13 See Crist (2005) for an in-depth discussion of such possible relations.

14 In our group discussions, we pondered the idea of also using `<entry>` elements for etymons, but dropped the idea in order not to introduce too much confusion to future users, and thinking that this would probably be harder to get into the TEI Guidelines later.

15 In the case of data sets (original or legacy) that do not use consistent terminology, variation in the terminology should be normalized to allow for maximally systematic search and retrieval possibilities.

16 `@part` does also have the option of the values `"Y"` yes and `"N"` no; however, given the options of initial/medial/final, yes and no are redundant and serve no additional value. We therefore do not recommend them.

17 https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#entries.

18 https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#cross-references.

19 Note that `@prev` and `@next` can be used to denote temporal sequences of referenced forms.

AUTHORS

JACK BOWERS

Jack Bowers is a linguist and language technologist with a wide array of interests, including: language documentation, cognitive semantics, diachronic linguistics, digital humanities and corpus linguistics. He received his M.A. in Linguistics from San José State University (San José, California) in 2012. He received his Ph.D. from the Université PSL at the École Pratique des Hauts Études (Paris, France) in 2020 for his dissertation Language Documentation and Standards in Digital Humanities: TEI and the documentation of Mixtepec-Mixtec. Jack has been active in the TEI community since 2014, particularly in the areas of lexical data, annotation, and especially, machine readable dictionaries. He is an Expert in the International Standards Organization (ISO) Technical Committee for Language and Terminology (TC37). He has worked at: Inria (Paris, France), the Austrian Center for Digital Humanities (Vienna, Austria), and currently works at Amazon Web Services (AWS) in the San Francisco Bay Area (USA).

AXEL HEROLD

Axel Herold is a researcher at the Berlin-Brandenburg Academy of Sciences and Humanities. His research centers on the modeling of lexicographic data of digitized as well as born-digital resources. He has worked on several European and German projects such as the German branch of the CLARIN European infrastructure focusing on the reuse and interoperability of lexical resources. He is the coordinator for the lexical resources task area in the German Text+ project (https://www.text-plus.org/).

TOMA TASOVAC

Toma Tasovac is Director of the Belgrade Center for Digital Humanities (BCDH) and Director of the Digital Research Infrastructure for the Arts and Humanities (DARIAH). His areas of interest include lexicography, data modeling, TEI, digital editions and research infrastructures. He is the co-leader of the DARIAH Working Group on Lexical Resources and co-author of TEI Lex-0, a baseline encoding for lexicographic data, which received the 2020 Rahtz Prize for TEI Ingenuity.

LAURENT ROMARY

Laurent Romary is Director for scientific information and Culture at Inria, France, and former initiator and director general of the DARIAH European infrastructure. He carries out research on the modeling of semi-structured documents, with a specific emphasis on texts and linguistic resources. He has been active in standardization activities with ISO, as chair of the ISO/TC 37/SC 4 (2002–2014) and ISO/TC 37 (2016–) committees, and with the Text Encoding Initiative, as member (2001–2011) and chair (2008–2011)

of its technical council. He has been involved in scientific information (now called open science) policies and corresponding infrastructure deployments (HAL and Episciences) since 2005 within various research-performing organizations.