eLex 2019,
Sintra, Portugal

TEI Lex-0 in action:

Improving the Encoding of the Dictionary of the *Academia das Ciências de Lisboa*

Ana Salgado[1], **Rute Costa**[1], **Toma Tasovac**[2], **Alberto Simões**[3]

[1] NOVA CLUNL, Universidade NOVA de Lisboa

[2] Belgrade Center for Digital Humanities, Serbia

[3] 2Ai – Instituto Politécnico do Cávado e do Ave / Algoritmi, Universidade do Minho

anasalgado@campus.fcsh.unl.pt; rute.costa@fcsh.unl.pt; ttasovac@humanistika.org; asimoes@ipca.pt

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

TEI Lex-0 in action

elexis european lexicographic infrastructure

NOVAFCSH

CLUNL

FCT Fundação para a Ciência e a Tecnologia

This research has been supported 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/03213/2019

- European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Overview

1. Goals
2. Dictionary of the *Academia das Ciências de Lisboa* (DACL)
3. TEI Guidelines for Dictionary encoding
4. TEI Lex-0 encoding of the DACL
5. TEI encoding of different types of lexical items
6. Conclusions and future work

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

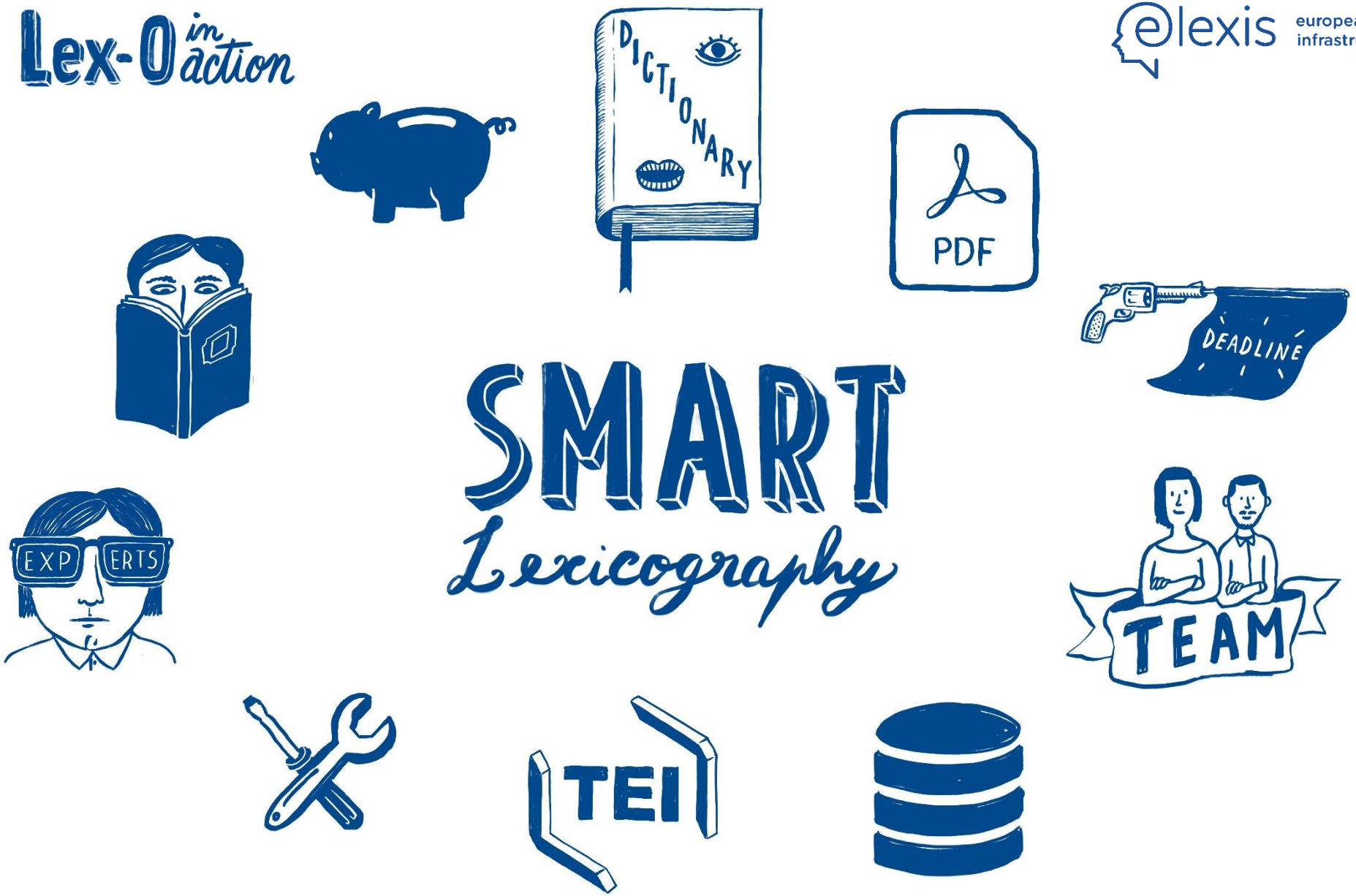
1. Goals

- to describe some experiments made while encoding the first complete dictionary of the *Academia das Ciências de Lisboa* (DACL)
- to discuss the TEI Lex-0 encoding
- to contribute to the efforts of the TEI Lex-0 group (DARIAH-ERIC Lexical Resources group)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

TEI Lex-0 in action

elexis european lexicographic infrastructure



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

2. Portuguese Academy Dictionary

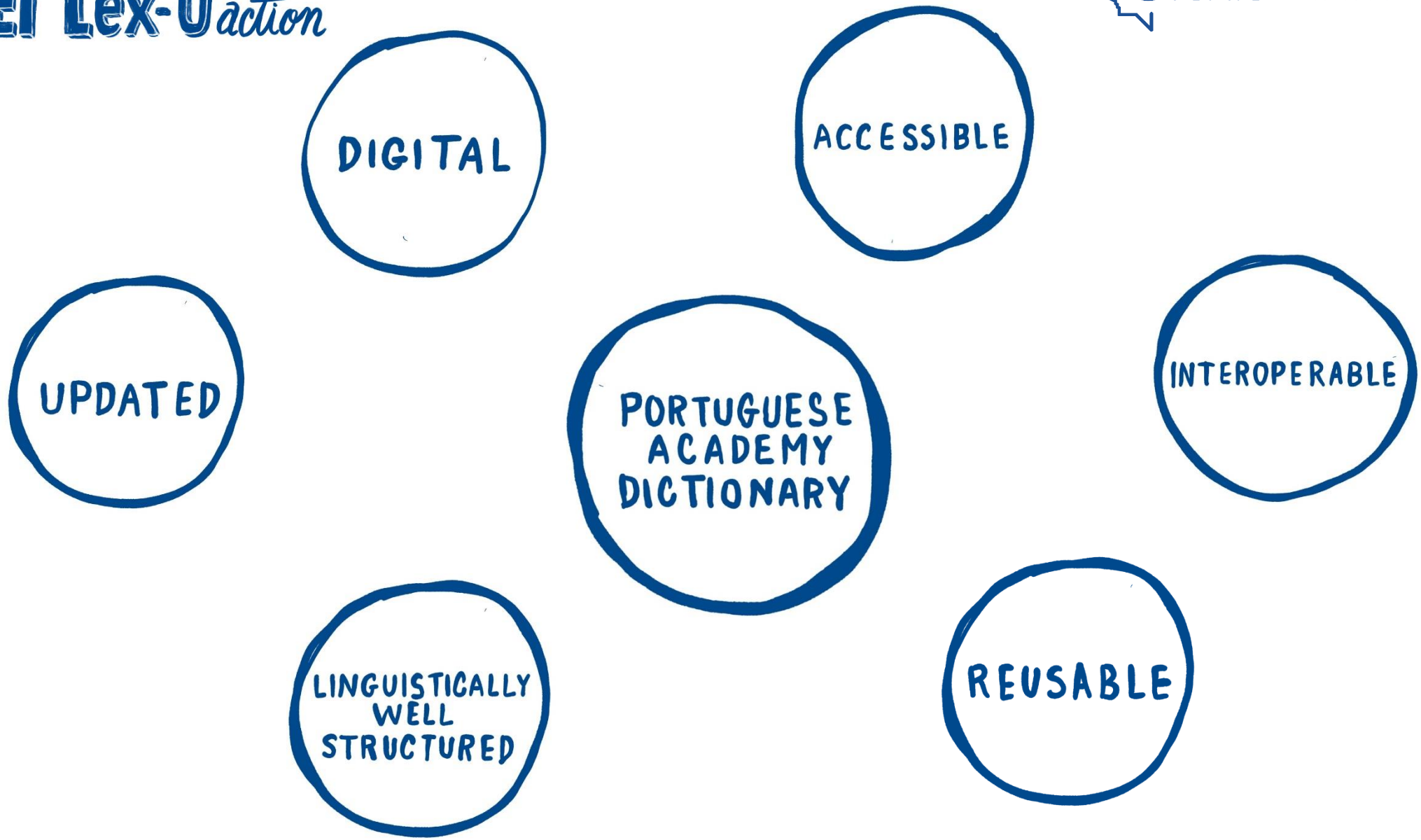
DLPC = *Dicionário da Língua Portuguesa Contemporânea* (2001)

69 426 entries, 167 556 senses

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

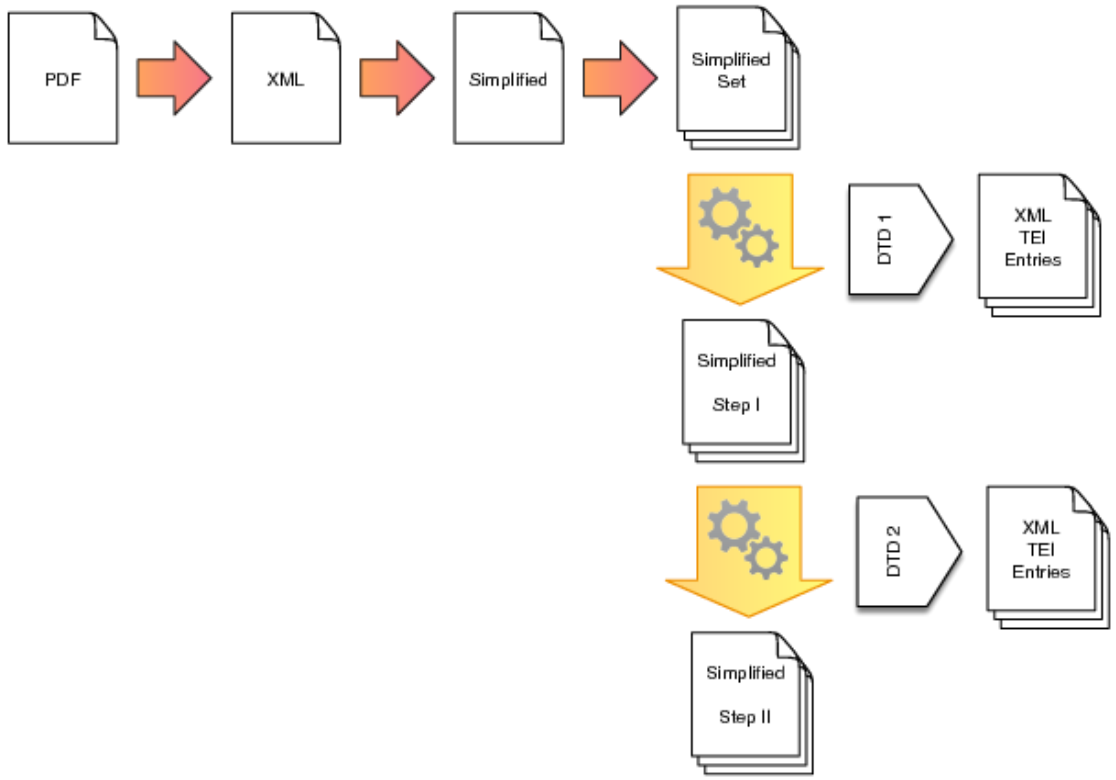
TEI Lex-0 in action

elexis european lexicographic infrastructure



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

2. Portuguese Academy Dictionary

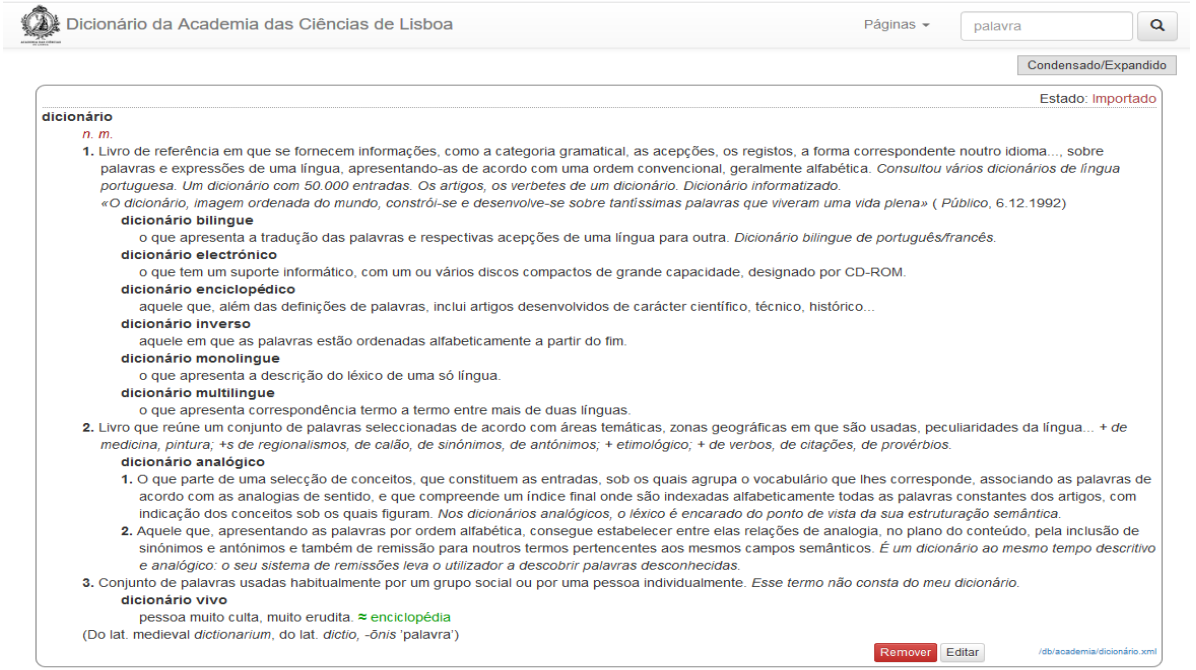


Simões et al. (2016). Building a Dictionary using XML Technology

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

2. Portuguese Academy Dictionary

http://www.lexmart.eu/



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

3. TEI Guidelines for Dictionary encoding

TEI dictionary encoding

- TEI is a *de facto* standard in digital edition or text annotation projects, from novels, letters to music notation.
- TEI has a specific module for encoding dictionaries.

BUT…

- We could not find solutions in the Guidelines that covered all the microstructural elements of the dictionary.
- TEI Guidelines contain multiple encoding possibilities for the same dictionary features.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

3. TEI Guidelines for Dictionary encoding

a[5] [ɐ]. *prep.* (Do lat. *ad* 'para' ou *ab* 'de'). **A** Valores semânticos: **I.** Na expressão de valores locativos, indica: **1.** Direcção para um lugar (real ou virtual). *O navio rumou a oriente. Levar a uma situação embaraçosa. Foi a casa dos sogros. Eu apenas fui a Paris; o meu irmão é que foi para Paris.* Obs. Quando introduz um complemento do verbo *ir* ou do nome *ida*, indica que a permanência no lugar de destino é breve; inversamente o uso da preposição *para* indica permanência prolongada. **2.** Termo de um movimento. *Chegou a casa.* **3.** Afastamento. ≃ DE. *Esquivar-se a trabalhos.* **4.** Distância medida em unidades de espaço ou tempo. *Há uma estação de comboio a quinhentos metros daqui. A minha casa fica a cinco minutos do mercado.* **5.** Localização, situação precisa ou aproximada. *Ela mora num palacete a São Bento. Pôs as cadeiras a todo o comprimento da sala.* **6.** Adjunção. *Amarrou o cão a um poste. A uma asneira seguiu-se outra.* **II.** Na expressão de valores temporais, indica: **1.** Tempo em que uma coisa acontece (pontual ou habitualmente); concomitância. *Tenho aulas a meio da tarde.* **2.** Distância. *O jogo está a dez minutos do intervalo. A cinco horas do desembarque.* **3.** Progressão para um tempo (em correlação com a prep. *de*). *De mês a mês. De cinco dias a esta parte. A exposição estará aberta ao público de Junho a Setembro.* **4.** Intervalo regular ou duração periódica. *Ele trabalhava a tempo inteiro. Há muitos contratos a prazo.* **III.** Na expressão de outros valores, indica: **1.** Causa. ≃ POR. *Fez isso a solicitação dos parentes.* **2.** Instrumento, meio e modo. *Pintura a óleo. Navegava a todo o vapor. O móvel apresentava entalhaduras a canivete. Há quem aguente muito tempo a pão e água.* **3.** Finalidade. ≃ PARA. *O patrão deu-lhe vinho a beber. Pôs*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

4. TEI Lex-0 encoding of the DACL

TEI Lex-0

Romary & Tasovac (2018). TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources

- a streamlined version of the TEI Guidelines
- given its (still) non-standard nature, it can be changed in order to accommodate relevant dictionary structures
- we have been participating in the TEI Lex-0 discussion: https://github.com/DARIAH-ERIC/lexicalresources/tree/master/Schemas/TEILex0
- work in progress

TEI Lex-0 will not replace the Dictionaries chapter in the TEI Guidelines.

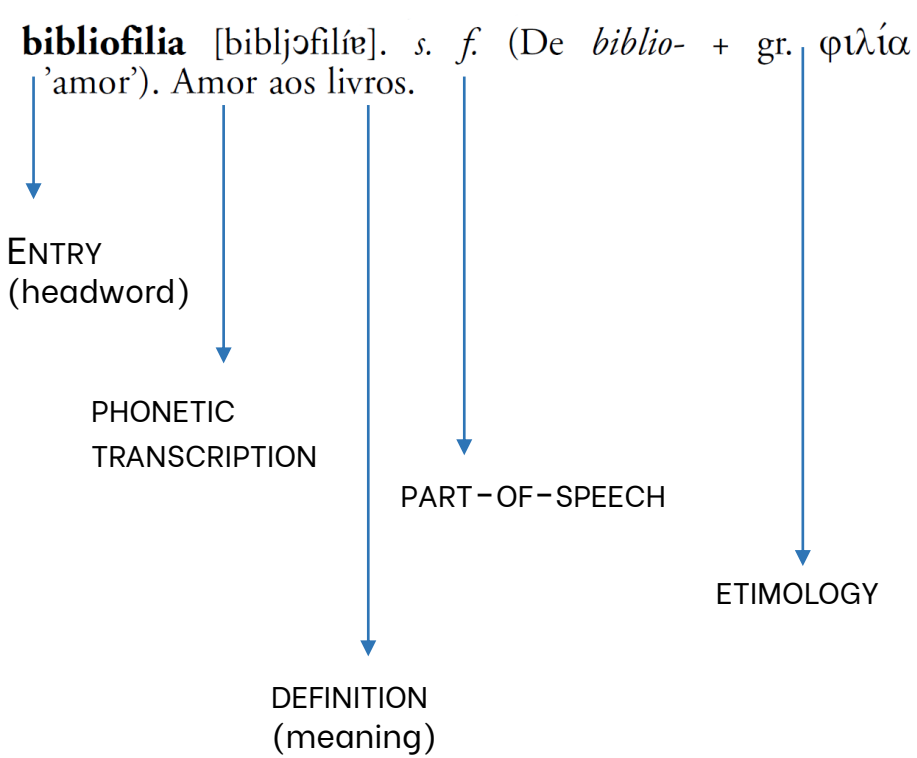
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

4. TEI Lex-0 encoding of the DACL

XML ESSENTIAL CHANGINGS (some examples)	
Original encoding TEI P5	Conversion into TEI Lex-0
<entry>	<entry xml:id="|*id*|" xml:lang="pt">
<term>	<form type="lemma">
<gramGrp>|*part of speech* and *gender*|</gramGrp>	<gramGrp><gram type="pos"></gram> <Gram type="gen">f.</gram></gramGrp>
<sense>	<sense xml:id="|*id*|">
<quote type="example">	<cit type="example"><quote>
<syn>|*synonym*|</syn>	<xr type="synonym"><ref type="entry|sense">|*synonym*|</ref></xr>
<cit><quote>|*example*|</quote> <bibl>|author|,<title>|title|</title>,|page|</bibl></cit>	<cit type="example"><quote>|*example*|</quote> <bibl><author>|*author*|</author> <title>|*title*|</title><citedRange>|*page*|</citedRange></bibl></cit>
<sense><def>V.<xr><ref>|*cross-reference*|</ref></xr></def></sense>	<xr><lbl>V.</lbl><ref>|*cross-reference*|</ref></xr>

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Basic entry structure



bibliofilia [bibliophilia], DLPC (2001)

```
<entry xml:lang="pt" xml:id="bibliofilia">
    <form type="lemma">
      <orth>bibliofilia</orth>
      <pron>bibljɔfilˈiɐ</pron>
    </form>
    <gramGrp>
      <gram type="lexicalConstruction"value=monolexical></gram>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">f.</gram>
    </gramGrp>
    <etym type="inheritance">
      <lbl>De</lbl>
      <cit type="etymon">
      <form>
        <orth>biblio-</orth>
      </form>
<etym type="grammaticalization">
          <seg type="desc">De</seg>
          <cit type="etymon">
            <form>
              <orth extent="pref">biblio-</orth>
            </form>
          </cit> <lbl>+</lbl>
<etym type="inheritance">
          <seg type="desc">Do</seg>
          <cit type="etymon" xml:lang="grc">
            <form><orth>φιλία</orth></form>
          </cit>
        </etym>
      <sense>
      <def>Amor aos livros</def><pc>.</pc>
      </sense>
</entry>
```

entry
- @xml:id
- @xml:lang – code (BCP 47)

gram type element
- part-of-speech of the entry and further specifications
- @norm attribute – values from the Universal Dependencies Part-of-Speech: https://universaldependencies.org/u/pos/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

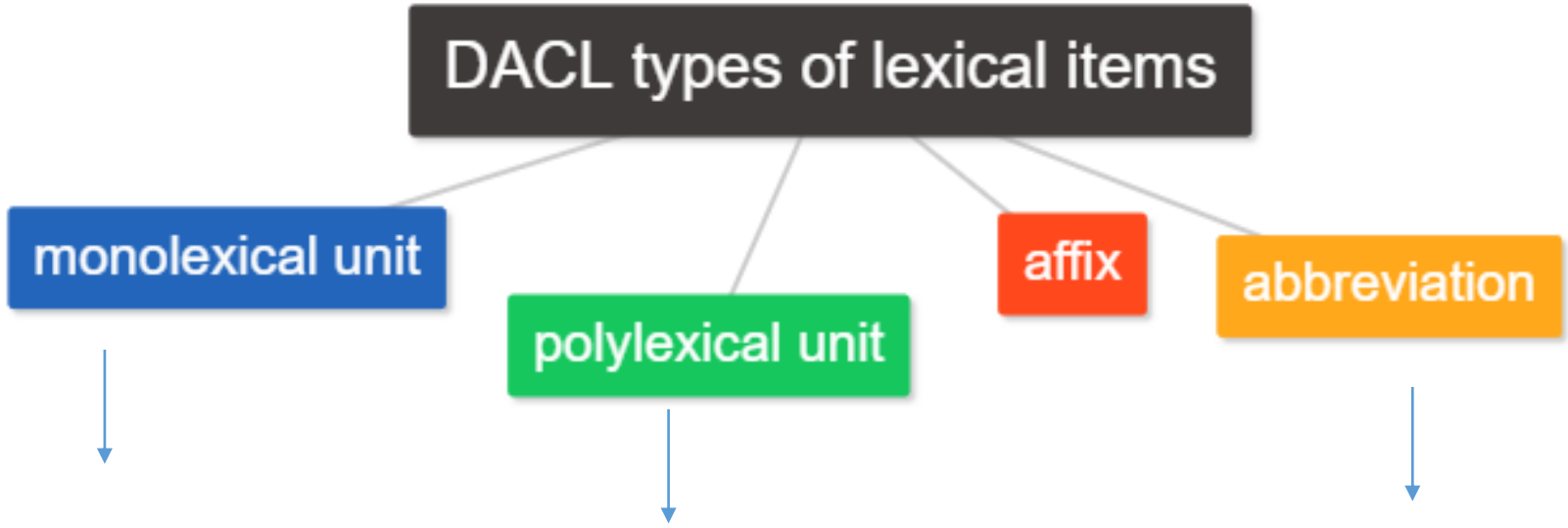
4. TEI Lex-0 encoding of the DACL

TEI Lex-0 classification of entries

- mainEntry
- relatedEntry
- homonymicEntry
- wordFamily

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

5. TEI encoding of different types of lexical items



lexical units, such as nouns, adjectives, verbs

grammatical units, such as conjunctions, determiners, prepositions, and pronouns

compounds and all kinds of lexical combinations, such as collocations or phrasemes

abbreviation (*Cf.*), alphabetism (*AAC*), acronym (*VIP*), symbol (*Ag*), contractions (*do* [of the]) and clipped forms (*metro* [metropolitan])

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Monolexical unit

palácio [pɐlásju]. *s. m.* (Do lat. *palatium*). **1.** Edifício sumptuoso, de grandes dimensões, geralmente construído num espaço urbano e destinado a residência da família real, de personalidades nobilitadas, de dignidades eclesiásticas ou altas individualidades. + *ducal, episcopal, presidencial, real.* **2.** Edifício sumptuoso, de dimensões significativas, onde se encontram sediados determinados organismos públicos. **palácio da justiça,** edifício, em cada localidade, onde funcionam os serviços judiciais e se realizam os julgamentos. *Um advogado seu amigo trabalha no palácio da justiça.* **3.** Casa solarenga, ampla, sumptuosa que lembra um palácio. **olhar para** *alguma coisa* **como boi para palácio.** 1. Não perceber nada de alguma coisa. 2. Não ligar importância; não dar valor, apreço a. Dim. palacete.

Original encoding

```
<entry id="palácio">
  <form>
    <orth>palácio</orth>
    <pron>pɐlˈasju</pron>
  </form>
  <gramGrp>s. m.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="main entry" xml:lang="pt"
xml:id="palácio">
  <form type="lemma">
     <orth>palácio</orth>
     <pron>pɐlˈasju</pron>
   </form>
  <gramGrp>
     <gram type="lexicalConstruction"
value="monolexical"></gram>
        <gram type="pos"
norm="NOUN">s.</gram>
        <gram type="gen">m.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

TEI Lex-0 in action

elexis european lexicographic infrastructure

Monolexical unit

workshop [wórkʃɔp]. *s. m.* (Ingl.). Reunião destinada à discussão ou realização de trabalho prático sobre um assunto específico, em que é feita uma aprendizagem através da troca de conhecimentos e experiências. *«Durante o 'workshop' sobre a articulação dos hospitais com os tribunais, foi visível a desconfiança de algumas pessoas»* (*DN*, 21.2.1992). Pl. workshops.

Original encoding

```
<entry id="workshop">
  <form>
   <orth>workshop</orth>
    <pron>wˈorkʃɔp</pron>
  </form>
  <gramGrp>s. m.</gramGrp>
  <etym>Ing.</etym>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="main entry" xml:lang="en"
xml:id="workshop">
  <form type="lemma">
    <orth>workshop</orth>
    <pron>wˈorkʃɔp</pron>
  </form>
  <gramGrp>
<gram type="lexicalConstruction"
value="monolexical"></gram>
     <gram type="pos" norm="NOUN">s.</gram>
     <gram type="gen">m.</gram>
  </gramGrp>
  <etym
type="borrowing"><lang>Ing.</lang></etym>
<!--etc. -->
</entry>
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

TEI Lex-0 in action

Monolexical unit

ensonado, a [ẽsunádu, -ɐ]. *adj.* (De *en-* + *sono* + suf. *-ado*). Que tem ou está com sono. ≃ SONOLENTO. *«Sertório assoma à porta do quarto: vem, ensonado, a esfregar os olhos.»* (D. MOURÃO-FERREIRA, *Gaivotas em Terra*, p. 139).

Original encoding

```
<entry id="ensonado">
  <form>
    <orth fem="a">ensonado</orth>
    <pron>ẽsunˈadu, -ɐ</pron>
  </form>
  <gramGrp>adj.</gramGrp>
<!--etc. -->
</entry>
```

elexis european lexicographic infrastructure

Conversion to TEI Lex-0

```
<entry type="main entry" xml:lang="pt"
xml:id="ensonado">
  <form type="lemma">
    <orth>ensonado</orth>
      <gramGrp>
     <gram type="lexicalConstruction"
value="monolexical"></gram>
      </gramGrp>
  </form>
  <form type="inflected">
    <orth>ensonado</orth>
    <pron>ẽsunˈadu</pron>
  <gramGrp>
     <gram type="gen">m.</gram>
  </gramGrp>
</form>
  <form type="inflected">
    <orth>ensonada</orth>
    <pron>ẽsunˈadɐ</pron>
  <gramGrp>
<gram type="gen">f.</gram>
  </gramGrp>
</form>
  <gramGrp>
     <gram type="pos" norm="ADJ">adj.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

TEI Lex-0 in action

elexis european lexicographic infrastructure

Polylexical unit

decreto-lei [dɨkrɛtulɐ̟j]. *s. m. Dir.* Acto normativo proveniente do Governo da República. *Actualmente, os decretos-leis são publicados na primeira série-A do Diário da República.* Pl. decretos-leis.

Original encoding

```
<entry id="decreto-lei">
  <form>
    <orth>decreto-lei</orth>
    <pron>dɨkrɛtulˈɐ̟j</pron>
  </form>
  <gramGrp>s. m.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="main entry" xml:lang="pt"
xml:id="decreto-lei">
  <form type="lemma">
    <orth>decreto-lei</orth>
    <pron>dɨkrɛtulˈɐ̟j</pron>
  </form>
  <gramGrp>
<gram type="lexicalConstruction"
value="polylexical"></gram>
    <gram type="pos" norm="NOUN">s.</gram>
    <gram type="gen">m.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

TEI Lex-0 in action

elexis european lexicographic infrastructure

Affix

(-)carpo(-) *elem. de form.* (Do gr. καρπός 'fruto'). Exprime a noção de *fruto*. *Mesocarpo, carpologia, pericarpo.*

Original encoding

```
<entry id="carpo">
  <form>
    <orth>(-)carpo(-)</orth>
  </form>
  <gramGrp>elem. de form.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="main entry" xml:lang="pt"
xml:id="carpo">
  <form type="lemma">
    <lbl>(-)</lbl><orth>carpo</orth><lbl>(-)</lbl>
  </form>
  <gramGrp>
<gram type="lexicalConstruction"
value="affix"></gram>
     <gram type="pos">elem. de form.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

TEI Lex-0 in action

elexis european lexicographic infrastructure

Abbreviation

VIP [víp]. *s. m. e. f.* Sigla de *Very Important Person* (Pessoa Muito Importante).

Original encoding

```
<entry id="VIP">
  <form>
    <orth>VIP</orth>
    <pron>víp</pron>
  </form>
  <gramGrp>s. m. e. f.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="main entry" xml:lang="pt"
xml:id="VIP">
  <form type="lemma">
    <orth>VIP</orth>
    <pron>víp</pron>
  </form>
  <gramGrp>
    <gram type="lexicalConstruction"
value="abbreviation"></gram>
     <gram type="pos" norm="NOUN">s.</gram>
     <gram type="gen">m.</gram>
     <lbl>e</lbl>
<gram type="gen">f.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

TEI Lex-0 in action

elexis european lexicographic infrastructure

Abbreviation

metro[2] [mɛ́tru]. *s. m.* (Red. de *metropolitano*). **1.** Sistema de transporte urbano efectuado por comboios de tracção eléctrica, em linhas parcial ou totalmente subterrâneas. ≃ METROPOLITANO. *Encontraram-se na estação de metro. O metro está em greve.* **boca**[+] **de metro. 2.** Comboio que assegura esse sistema de transporte. *Apanhar, perder o +.*

Original encoding

```
<entry id="metro:2">
  <form>
    <orth>metro:2</orth>
    <pron>mˈɛtru</pron>
  </form>
  <gramGrp>s. m.</pos>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="homonymicEntry" xml:lang="pt"
xml:id="metro_2 n="2">
  <form type="lemma">
    <orth>metro</orth>
    <pron>mˈɛtru</pron>
  </form>
  <gramGrp>
    <gram type="lexicalConstruction"
value="abbreviation"></gram>
     <gram type="pos" norm="NOUN">s.</gram>
     <gram type="gen">m.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

TEI Lex-0 in action

elexis european lexicographic infrastructure

Conclusions and future work

- This task has made it possible to evaluate the consistency of the data referring to the DACL print edition, thereby highlighting content-related inconsistencies and, ultimately, contributing towards the optimization of lexicographic teamwork.
- This is a work in progress and further discussion on how to encode most of the information properly is still needed.
- TEI Lex-0 is stricter than TEI, but it is fully capable of representing the complexities of the entry structure of the DACL.
- Sometimes the encoding is more verbose but more structural, allowing machines to process this information better.
- Our present goal is not to have the dictionary in TEI Lex-0 only.
- An agreement between academies and other institutions would be desirable to systematize and optimize resources that can provide a better representation of the entire European lexicographical heritage.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

TEI Lex-0 in action

elexis european lexicographic infrastructure

Thank you for your attention.

Obrigada pela vossa atenção.

anasalgado@campus.fcsh.unl.pt
rute.costa@fcsh.unl.pt
ttasovac@humanistika.org
asimoes@ipca.pt

NOVAFCSH Faculdade de Ciências Sociais e Humanas Universidade Nova de Lisboa

CLUNL Centro de Linguística da Universidade Nova de Lisboa

BELGRADE CENTER FOR DIGITAL HUMANITIES

2Ai APPLIED ARTIFICIAL INTELLIGENCE LABORATORY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.