



# Le projet MorDigital et l'emploi de TEI Lex-0

Rute Costa & Ana Salgado

NOVA FCSH, Lisbonne

8 février 2023



# Acknowledgments

- MORDigital – *Digitalização do Diccionario da Lingua Portugueza de António de Morais Silva* [PTDC/LLT-LIN/6841/2020] project financed by the Portuguese National Funding agency through the FCT – Fundação para a Ciência e Tecnologia
- 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

# Outline

- Introduction
- Historical background
- Morais dictionary
- MORDigital project
- Following best practices: using TEI Lex-0
- Concluding remarks

# Introduction

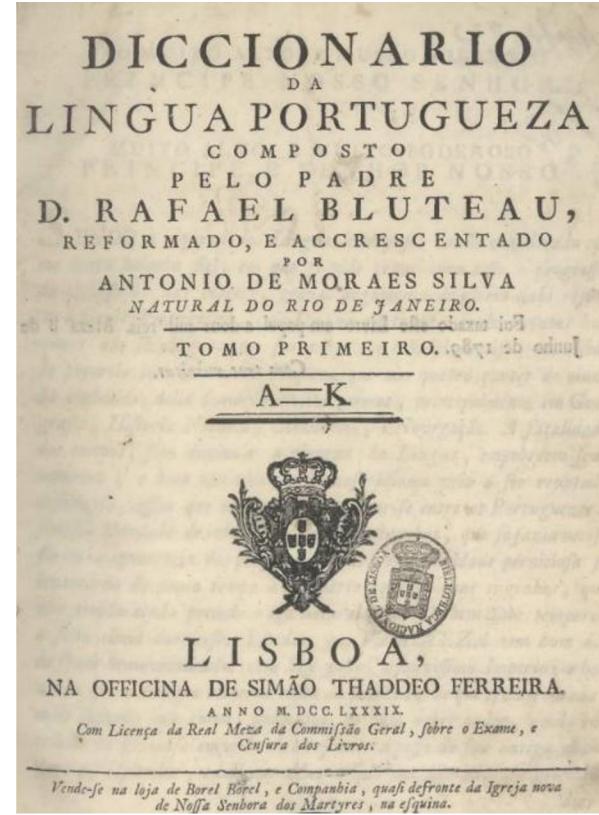
- The MORDigital project started in March 2021 and will be founded till next year.
- The project aims to produce high-quality and searchable digital versions of the first three editions (1789; 1813; 1823) of Morais dictionary in order to preserve this important European heritage work.
- This project aims to contribute substantially to the scientific community and aspires to apply innovative computational methodologies.

# Historical background

- *Diccionario da Lingua Portugueza* by António de Morais Silva was elaborated during the Age of Enlightenment.
- The eighteenth century brought a renewal in several fields of knowledge, namely those concerning the description of living languages, at a time when Latin was still the language of instruction.
- The publication of the Morais dictionary in 1789 inaugurated modern Portuguese lexicography.

# Morais dictionary

- The *Diccionario da Lingua Portugueza* by António de Morais Silva marks the beginning of modern Portuguese lexicography and serves as a model for all subsequent lexicographic production throughout the 19th and 20th centuries.
- Morais does not claim to be the author, assigning this condition to Bluteau, author of the *Vocabulario Portuguez and Latino*.
- Morais recognises in the ‘*Prólogo ao Leitor*’ [Prologue to the Reader] that the additions he brought to the dictionary are quite relevant.



Frontispiece of Morais (1789), first volume

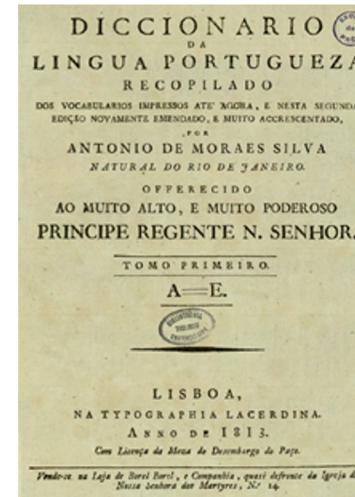
# Moraïs dictionary

## The *Diccionario da Lingua Portugueza*

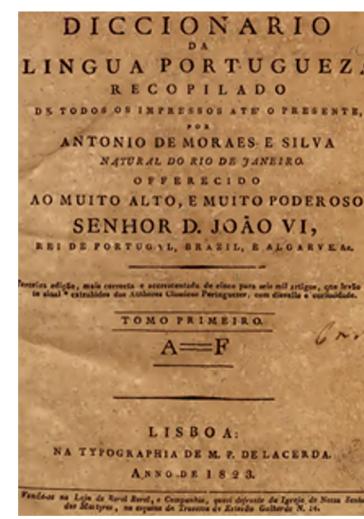
Frontispiece of Moraïs dictionaries (1789, 1813, 1823)



Two volumes  
A to K, 752 pp.  
L to Z, 541 pp.



Two volumes  
A to E, 889 pp.  
F to Z, 886 pp.



Two volumes  
A to K, 952 pp.  
L to Z, 872 pp.

# MORDigital project

<https://mordigital.fcsh.unl.pt>



# MORDigital project

- Although it is a Portuguese national project, we are an international team with different backgrounds, terminology, lexicography, linguistic linked data, computer science and digital humanities.
- The project also aims to show the advantages of structured digital versions of dictionaries in combining lexicographic methodologies with terminological methods.



<https://mordigital.fcsh.unl.pt/en/team/>

# MORDigital project

- NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal
- Academia das Ciências de Lisboa, Portugal
- Istituto Di Linguistica Computazionale ‘A. Zampolli’, Italy
- CLLC, Centro de Línguas, Literaturas e Culturas da Universidade de Aveiro, Portugal
- Inria, team ALMAnaCH, France
- ROSSIO Infrastructure, Portugal
- Arcascience, France
- BCDH – Belgrade Center for Digital Humanities, Serbia



*Inria*



# MORDigital project

## MORDigital aims

- to encode the selected editions of *Diccionario de Lingua Portugueza* by António de Moraes Silva
- to promote accessibility to cultural heritage while fostering reusability and contributing towards a greater presence of lexicographic digital content in Portuguese through open tools and standards
- to connect data and metadata within the same lexicographic resource and between different resources, through the Web of Data

These digital versions will be converted into structured data.

This dictionary will also be made available via an online interface on the website (at the moment only PDFs are available).

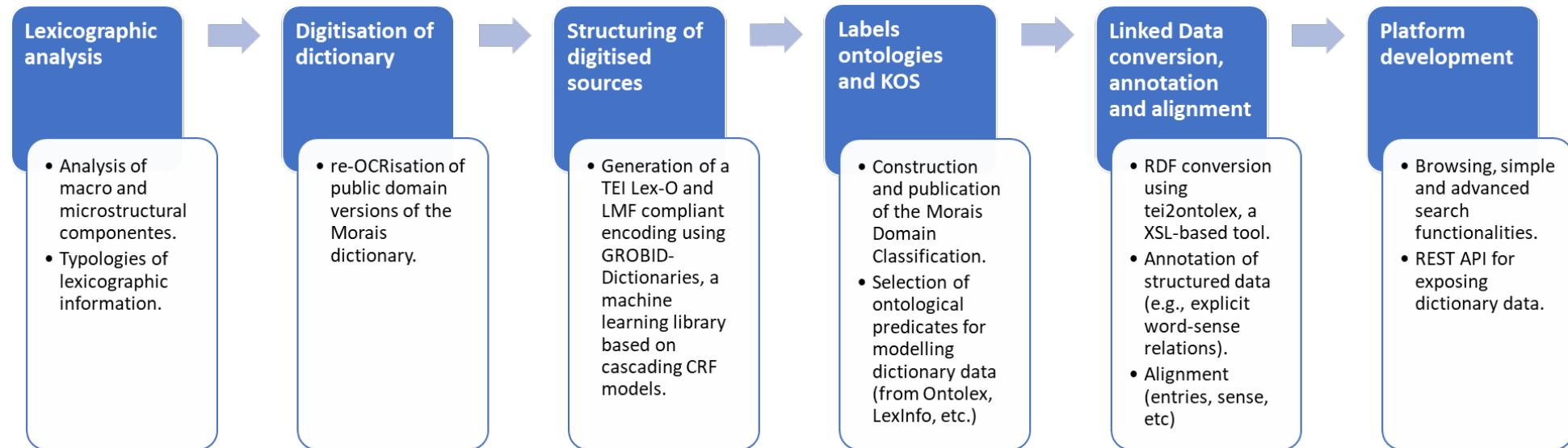


<https://mordigital.fcsh.unl.pt>

## Methodology

- To analyse all components that comprise the dictionary's macro- and microstructure;
- To identify, organise and describe the different levels of linguistic knowledge to apply the aforementioned standards systematically;
- To develop methodologies that can be replicated for other applications and test the alignment of the different encodings of Morais;
- To participate in reviewing the corresponding standards as members of the standard bodies and scientific forums;
- To propose best practices for harmonising the encoding of lexicographic resources;
- To make Morais available via an open-access platform.

## Workflow



# MORDigital project

# Optical Character Recognition (OCR) Corrections

- We already had access to OCR'ed versions of the dictionary editions at the beginning of the project.
  - These needed to be post-corrected. For this we decided to use **ABBYY FineReader**.
  - An established tool with user-friendly interface.
  - It creates editable, searchable files.  
Produces several output formats (RTF, DOCX, PDF, HTML, XML, etc.).
  - The tool preserves typographical features.



# MORDigital project

Noise	Type	Action: insert	Action: replace	Frequency per page (sample)
\$	symbol	§		A37 (3)
%	symbol	§		A37 (7)
v. g.	1st character non-italic		v. g.	
i; /; f; í	instead of long /s/	f		
íl; jj; (T;	instead of double long /s/	ll		
alguém	graphic accent / ' /		alguem	A75 (5); A76 (2); A77 (2); A80 (4); A88 (1); A89 (2); A91 (4); A95 (2); A96 (2)
á	graphic accent / ' /		ã	A79 (10), A80 (3), A81 (13), A76 (6), A77 (8), A78 (5), A85 (4), A91 (4), A92 (6), A93 (8), A94 (11)
Adadeira	the 1st letter of most capitalised forms in italic is replaced by 2 letters, e.g., /Ad/ instead of /M/		Madeira	
Fieira	the 1st letter of a specific proper noun in italic is recursively replaced, i.e., /F/ instead of /V/		Vieira	
d; if; fl; iT;	ligature e.g. [activamente]:  activamente.		ct	
r	typographic		t	

Some common recurrent types of noise caused by the optical character recognition tool

## Optical Character Recognition (OCR) Corrections

Original text	Re-OCRization before cleaning the OCR noise	Print after cleaning the OCR noise
<p>PASTOR, f. m. o que guarda , e apascenta o gado. § f. Pastor, o Cura d'almas , e todo o ministro da Igreja , que administra o pasto espiritual. § O Rei como diz Homero deve fer pastor do seu povo , i. e. administrar-lhe de que viva farto , defende lo dos inimigos internos , e externos ; e tirar delle só o que bastrar para as necessidades suas , e do público. Barros. Elogio 1.</p>	<p>PASTOR, ri m. o que guarda , e apascenta o gado. § ri Pastor, o Cura d'almas , e todo o Wlinistro da Igreja , que administra o pasto epi-ritual. § O Rei como tfa, Homero deve fer paftor do feu povo , i. e. adtrimi ftrar-lhc, de qué viva farto , defendçlo dos inimigos internos., e ex-ternos ; e tirar delle só o que baftar para as necelfidadési, suas , e do público. Barras. -Elo-gio 1.</p>	<p>PASTOR , f. m. o que guarda , e apascenta o gado. § f. Pastor , o Cura d'almas , e todo o ministro da Igreja , que administra o pasto espiritu-al. § O Rei como diz Homero deve fer pastor do seu povo , i. e. administrar-lhe de que viva farto , defende lo dos inimigos internos , e exter-nos ; e tirar delle só o que bastrar para as necessidades suas , e do público. Barros. Elogio 1.</p>

Lexicographic article of PASTOR before and after cleaning the OCR noise

# MORDigital project

- For the structuring of the digital editions we are using **GROBID-Dictionaries**.
- This is a machine learning library for structuring digitised lexical resources and entry-based documents with encyclopedic or bibliographic content.
- In particular, it allows the automatic parsing, extraction and structuring of lexical information from PDFs.
- GROBID-Dictionaries takes as input lexical resources digitised in PDF format and generates a TEI-encoded hierarchy of the different text structures which it has recognised.

# Following best practices

- There are two initiatives which have been adopted in general:
  - 1) the **TEI Guidelines** and its specific module for dictionaries in Chapter 9 ('Dictionaries') by TEI Consortium or **TEI Lex-0**, a technical specification and a set of community-based recommendations for encoding machine-readable dictionaries hosted by the DARIAH Working Group Lexical Resources;
  - 2) **the Lexicon Model for Ontologies (Ontolex-Lemon)**, together with the Lexicography Module (lexicog) from the Ontolex Lexicon Community group.
- Our aim is to convert Morais dictionary into a structured lexical resource in both TEI-XML (based on the ISO LMF standard) and in RDF (based on the OntoLex-Lemon model and its recent extensions).
- The TEI-XML sources will subsequently be converted to OntoLex (both the original model and its follow-up modules) using an XSLT stylesheet.

# Following best practices

- The TEI header is a key element of the structure of any TEI document, in which the metadata of the encoded text is structurally stored, that is, where the detailed bibliographic data from both the printed source(s) and the electronic file are described in order to improve search engines.
  - The Morais dictionary TEI header consists of:
    - a file description;
    - an encoding description;
    - a profile description.

## TEI header: MORAIS dictionary (1st ed., 1789)

# Following best practices

```

<abbr type="POS" norm="adjective">adj.</abbr>
<abbr type="POS" norm="adverb">adv.</abbr>
<abbr type="POS" norm="conjunction">Conj.</abbr>
<abbr type="POS" norm="interjection">Interj.</abbr>
<abbr type="POS" norm="numeral">Num.</abbr>
<abbr type="POS" norm="adposition">Prep.</abbr>
<abbr type="POS" norm="pronoun">Pron.</abbr>
<abbr type="POS" norm="noun">S.</abbr>
<abbr type="POS" norm="verb">V.</abbr> <subc>at.</subc>
<abbr type="POS" norm="verb">V.</abbr> <subc>imperf.</subc>
<abbr type="POS" norm="verb">V.</abbr> <subc>n.</subc>
<abbr type="POS" norm="verb">V.</abbr> <subc>recipr.</subc>

<abbr type="geographic">Afiat.</abbr>

<abbr type="time">Ant.</abbr> || <abbr type="time">antiq.</abbr>

<abbr type="textType">Poet.</abbr>

<abbr type="socioCultural">Ch.</abbr> || <abbr type="socioCultural">Chul.</abbr>
<abbr type="socioCultural">Fam.</abbr>
<abbr type="socioCultural">Vulg.</abbr>

<abbr type="frequency">Freq.</abbr>
<abbr type="frequency">P. uſ.</abbr>

<abbr type="gender">Com.</abbr>
<abbr type="gender">F.</abbr>

<abbr type="number">Pl.</abbr>
<abbr type="number">Sing.</abbr>

```

```

<abbr type="domain">Agric.</abbr>
<abbr type="domain">Anat.</abbr>
<abbr type="domain">Archit.</abbr>
<abbr type="domain">Arithm.</abbr>
<abbr type="domain">Artelh.</abbr>
<abbr type="domain">Aſtrol.</abbr>
<abbr type="domain">Aſtron.</abbr>
<abbr type="domain">Botan.</abbr>
<abbr type="domain">Braſ.</abbr>
<abbr type="domain">Chim.</abbr>
<abbr type="domain">Cirurg.</abbr>
<abbr type="domain">Chron.</abbr> || <abbr type="domain">Cron.</abbr>
<abbr type="domain">Eſcult.</abbr>
<abbr type="domain">Filoz.</abbr>
<abbr type="domain">Fific.</abbr>
<abbr type="domain">Fortif.</abbr>
<abbr type="domain">Geogr.</abbr>
<abbr type="domain">Geometr.</abbr>
<abbr type="domain">Grammat.</abbr>
<abbr type="domain">Jurid.</abbr>
<abbr type="domain">Juriſp.</abbr>
<abbr type="domain">Log.</abbr>
<abbr type="domain">Manej.</abbr>
<abbr type="domain">Mathem.</abbr>
<abbr type="domain">Med.</abbr>
<abbr type="domain">Milit.</abbr>
<abbr type="domain">Muſ.</abbr>
<abbr type="domain">Naut.</abbr>
<abbr type="domain">Opt.</abbr>
<abbr type="domain">Ortogr.</abbr>
<abbr type="domain">Perſp.</abbr>
<abbr type="domain">Pharmac.</abbr>
<abbr type="domain">Pint.</abbr>
<abbr type="domain">Rhet.</abbr>
<abbr type="domain">Theol.</abbr>
<abbr type="domain">Volat.</abbr>

```

Abbreviations. Source: MORAIS dictionary (1st ed., 1789)

# Following best practices

- It is essential to use a consistent identification of content to improve its reusability, defining different levels of granularity.
- Concerning the `xml:id` attribute (whose value must be unique within a given XML document), we use a dot as a delimiter for all subsequent parts.
- The unique ids will be created automatically by an XSLT script.
- The id consists of the author's name, the edition number, the dictionary title abbreviated and a non-accented lemma, for example, "MORAIS.1.DLP.ABA".

# Following best practices

- We decided to keep the textual content exactly as it appeared in the printed edition.

ESTOJO , f. m. caixinha de coiro , ou papé-lão com repartimentos para navalhas , tesouras , facas , canivetes , &c.

```
<entry xmlns="http://www.tei-c.org/ns/1.0" xml:id="MORAIS.1.DLP.ESTOJO" type="mainEntry" xml:lang="pt">
  <form type="lemma">
    <orth>ESTOJO</orth>
  </form>
  <metamark function="lemmaDelimiter">,</metamark>
  <gramGrp>
    <gram type="pos" norm="NOUN">f.</gram>
    <gram type="gen">m.</gram>
  </gramGrp>
  <sense xml:id="MORAIS.1.DLP.ESTOJO.s.1">
    <def>caixinha de coiro , ou papélão com repartimentos para navalhas , tesouras , facas , canivetes , &amp;</def>
  </sense>
  <pc></pc>
</entry>
```

ESTOJO [case; cover; kit], example of a basic article structure.

# Following best practices

**JALDE ; adj. còr amarella acceza.**

```
<entry xmlns="http://www.tei-c.org/ns/1.0" xml:id="MORAIS.1.DLP.JALDE"
type="mainEntry" xml:lang="pt">
    <form type="lemma">
        <orth>JALDE</orth>
    </form>
    <metamark function="lemmaDelimiter">;</metamark>
    <gramGrp>
        <gram type="pos" norm="ADJECTIVE">adj.</gram>
    </gramGrp>
    <sense xml:id="MORAIS.1.DLP.JALDE.s.1">
        <def>còr amarella acceza</def>
    </sense>
    <pc>.</pc>
</entry>
```

JALDE [yellow color], an example an article with a semicolon delimiting the lemma from the POS.

**ABADERNÁS, plur. femin. naut. ganchos onde se fixão os colhedores, e outros cabos, quando se aperta a enxarcia.**

```
<entry xmlns="http://www.tei-c.org/ns/1.0" xml:id="MORAIS.1.DLP.ABADERNAS"
type="mainEntry" xml:lang="pt">
    <form type="lemma">
        <orth>ABADERNÁS</orth>
    </form>
    <metamark function="lemmaDelimiter">;</metamark>
    <gramGrp>
        <gram type="number">plur.</gram>
        <gram type="gender">femin.</gram>
    </gramGrp>
    <sense xml:id="MORAIS.1.DLP.ABADERNAS.s.1">
        <usg type="domain">naut.</usg>
        <def>ganchos onde se fixão os colhedores, e outros cabos, quando se aperta a
enxarcia</def>
    </sense>
    <pc>.</pc>
</entry>
```

ABADERNÁS

# Following best practices

ABADA, s. f. A porção que leva a aba colhida, e apanhada § n. prapr. de huma especie d'animal que tem ponta, e he o mesmo que *Rinoceronte*.

```
<entry xmlns="http://www.tei-c.org/ns/1.0" xml:id="MORAIS.1.DLP.ABADA" type="mainEntry" xml:lang="pt">
    <form type="lemma">
        <orth>ABADA</orth>
    </form>
    <gramGrp>
        <gram type="pos" norm="NOUN">f.</gram>
        <gram type="gen">f.</gram>
    </gramGrp>
    <sense xml:id="MORAIS.1.DLP.ABADA.s.1">
        <def>A porção que leva a aba colhida, e apanhada</def>
    </sense>
    <metamark function="senseDelimiter">§</metamark>
    <sense xml:id="MORAIS.1.DLP.ABADA.s.2">
        <def><hi rend="italic">n. prapr. de huma especie d'animal que tem ponta, e he o mesmo que <hi
            rend="italic">Rinoceronte</hi></def>
        <pc>.</pc>
    </sense>
</entry>
```

ABADA.

# Following best practices

## ABCESSO. v. abcesso.

```

<entry xmlns="http://www.tei-c.org/ns/1.0"
      xml:id="MORAIS.1.DLP.ABCESSO" type="mainEntry"
      xml:lang="pt">
  <form type="lemma">
    <orth>ABCESSO</orth>
  </form>
  <metamark function="lemmaDelimiter">.</metamark>
  <xr type="related">
    <lbl expand="veja"><hi>v.</hi></lbl>
    <ref target="#MORAIS.1.DLP.ABCESSO"
        type="mainEntry">abcesso</ref>
  </xr>
  <pc>.</pc>
</entry>
```

ABCESSO [abscess], cross-reference preceded by a *v.*

## ABADEJO, f. m. v. Vaca loura : v. *Badejo*:

```

<entry xmlns="http://www.tei-c.org/ns/1.0" xml:id="MORAIS.1.DLP.ABADEJO" type="mainEntry" xml:lang="pt">
  <form type="lemma">
    <orth>ABADEJO</orth>
  </form>
  <metamark function="lemmaDelimiter">,</metamark>
  <gramGrp>
    <gram type="pos" norm="NOUN">f.</gram>
    <gram type="gen">m.</gram>
  </gramGrp>
  <sense xml:id="MORAIS.1.DLP.ABADEJO.s.1">
    <xr type="related">
      <lbl expand="veja"><hi>v.</hi></lbl>
      <ref target="VACA-LOURA" xml:id="MORAIS.1.DLP.VACA-LOURA" type="mainEntry">Vaca loura</ref>
    </xr>
  </sense>
  <metamark function="senseDelimiter">:</metamark>
  <sense xml:id="MORAIS.1.DLP.ABADEJO.s.2">
    <xr type="related">
      <lbl expand="veja:"><hi>v.</hi></lbl>
      <ref target="BADEJO" xml:id="MORAIS.1.DLP.BADEJO" type="mainEntry"><hi>Badejo</hi></ref>
    </xr>
  </sense>
  <pc>.</pc>
</entry>
```

ABADEJO [stag beetle], cross-reference preceded by a *v.*, followed by a synonymous definition, a colon, *v.* and another cross-reference

# Following best practices

ABACO, f. m. Peça superior do capitel da columna, serve como de coberta ao cesto de flores, que nelle se representa ; usa-se na *Architect.* § t. *arithm.* a taboada de Pythagoras.

```
<entry xmlns="http://www.tei-c.org/ns/1.0" xml:id="MORAIS.1.DLP.ABACO"
type="mainEntry" xml:lang="pt">
  <form type="lemma">
    <orth>ABACO</orth>
  </form>
  <metamark function="lemmaDelimiter">,</metamark>
  <gramGrp>
    <gram type="pos" norm="NOUN">f.</gram>
    <gram type="gen">m.</gram>
  </gramGrp>
  <sense xml:id="MORAIS.1.DLP.ABACO.s.1">
    <!-- SEE usa-se na Architect. = domain -->
    <def>Peça superior do capitel da columna , serve como de coberta ao cesto de
flores , que nelle se representa ; usa-se na</def>
    <usg type="domain">Architect.</usg>
  </sense>
  <metamark function="senseDelimiter">§</metamark>
  <sense xml:id="MORAIS.1.DLP.ABACO.s.2">
    <usg type="domain">t. arithm.</usg>
    <def>a taboada de Pythagoras</def>
  </sense>
  <pc>.</pc>
</entry>
```

ABACO [abacus], example of a domain label inside the lexicographic definition.

# Following best practices

**ABAFADIÇO**, adj. v. g. *lugar* — calmofo , em que não corre o ar livremente , ou viração *B. Pereira*. § *F. homem*—que se afronta facilmente. *Uli-*  
*sípo* 262.

ABAFADIÇO [suffocating; airless; (person) irritable].

```

<entry xmlns="http://www.tei-c.org/ns/1.0" xml:id="MORAIS_1.DLP.ABAFADIÇO" type="mainEntry"
xml:lang="pt">
  <form type="lemma">
    <orth>ABAFADIÇO</orth>
  </form>
  <metamark function="lemmaDelimiter">,</metamark>
  <gramGrp>
    <gram type="pos" norm="ADJ">adj.</gram>
  </gramGrp>
  <sense xml:id="MORAIS_1.DLP.ABAFADIÇO.s.1">
    <lbl expand="verbi gratia" xml:lang="la"><hi>v. g.</hi></lbl>
    <cit type="example"><quote>lugar ——</quote></cit>
    <def>calmofo, em que não corre o ar livremente, ou viração</def>
  </sense>
  <metamark function="senseDelimiter">§</metamark>
  <sense xml:id="MORAIS_1.DLP.ABAFADIÇO.s.2">
    <cit type="example">
      <quote>homem ——</quote>
    </cit>
    <def>que se afronta facilmente</def>
    <pc>.</pc>
    <cit type="example">
      <bibl type="attestation">
        <!-- point to Uliſipo -->
        <title>Uliſipo</title>
        <citedRange unit="page">262</citedRange>
      </bibl>
    </cit>
  </sense>
</entry>
```

# Following best practices

FE'DO , adj. feio. *Luz da Medicina* „, lepra,  
e outros achaques fédos , p. usado.

```

<entry xmlns="http://www.tei-c.org/ns/1.0" xml:id="MORAIS1.DLP.FEDO" type="mainEntry" xml:lang="pt">
    <form type="lemma">
        <orth>FE'DO</orth>
    </form>
    <metamark function="lemmaDelimiter">,</metamark>
    <gramGrp>
        <gram type="pos" norm="ADJECTIVE">adj.</gram>
    </gramGrp>
    <sense xml:id="MORAIS1.DLP.FEDO.s.1">
        <def>feio</def>
        <metamark function="exampleDelimiter">,</metamark>
        <cit type="example" xml:lang="pt">
            <bibl type="attestation" source="#M._L._Monarchia_Lusitana">
                <title>Luz da Medicina</title>
            </bibl>
            <pc>,,</pc>
            <quote>lepra , e outros achaques fédos</quote>
        </cit>
        <metamark function="usageDelimiter">,</metamark>
        <usg type="frequency">p. usado</usg>
    </sense>
    <pc>.</pc>
</entry>
```

FEDO [ugly; nasty], an example of an article  
with a quote.

# Concluding remarks

- This project will contribute towards a more significant presence of lexicographic digital content in Portuguese through open tools and standards.
- The linking mechanisms of the resulting structured dictionary to other resources will constitute a prototype that can be replicated in other works, namely in the Portuguese-speaking world.
- We also propose combining semasiological and onomasiological approaches in our treatment of the different editions of Morais. For this, we foresee the inclusion of ontologies (e.g. diasystematic marking, namely domain labels, registers and part of speech categories).
- We expect to have encoded a vital heritage dictionary, compliant with the most advanced standards for scholarly digital editions and made available via an open license.

# Concluding remarks

- The versions will be accessible and searchable through an advanced interface, which will enable the selective querying of text by lemma and type of lexicographic content.
- The project will have significantly contributed towards the analysis and annotation of dictionaries through computer-assisted processes.
- Start OCR corrections on the two other editions of the dictionary
- Run the GROBID tool iteratively on the corrected output of the OCR of the first edition to ensure a correct TEI-XML (LMF) encoding of the different components of single entries (e.g., authoritative examples, collocations, )
- Start testing the XSLT transformation to OntoLex and its extensions
- In addition throughout the project our approach foresees a rigorous linguistic treatment of the source texts. This will make it possible to organise and structure the lexicographic components, and to elicit lexical relationships between various elements.
- We also propose combining semasiological and onomasiological approaches in our treatment of the different editions of Morais. For this we foresee the inclusion of ontologies (e.g. diasystematic marking, namely domain labels, registers and part of speech categories).

# Thank you!

## ... questions?

Rute Costa: [rute.costa@fcsh.unl.pt](mailto:rute.costa@fcsh.unl.pt)

Ana Salgado: [anasalgado@fcsh.unl.pt](mailto:anasalgado@fcsh.unl.pt)