

MorDigital



Morais dictionary: following best practices
in a retrodigitised dictionary project

*Ana Salgado, Laurent Romary, Rute Costa, Toma Tasovac, Anas Fahad Khan, Margarida Ramos,
Bruno Almeida, Sara Carvalho, Mohamed Khemakhem, Raquel Silva*



Outline

- Introduction
- MORDigital project
- Following best practices
 - 1) establishing the data model
 - 2) refining metadata
 - 3) using consistent identifiers
 - 4) enhancing the encoding
- Concluding remarks



What is a best practice?

- A best practice is a procedure that has been shown by research and experience to produce good outcomes and that is officially recognised as being effective.
- It may also be already established or have been proposed as a standard or set of guidelines for widespread adoption.
- For lexicographic works, adopting best practices has several advantages, especially for reasons of interoperability and also for supporting the longer-term sustainability of the content.



What is a retrodigisited dictionary project?

A lexicographic resource is a lexical resource that consists of a dataset that is human-readable as a dictionary and also can be processed as a machine-readable dictionary (MRD) model

– born-digital dictionaries, created as machine-readable

– retrodigisited dictionaries, which were converted from an analogue (paper) or digital (e.g., PDF) medium to a computer-readable format, using OCR systems and involving the encoding step of the scanned version

- Many institutions are now involved in mass digitisation projects to make historical documents available online.

MORDigital project

- MORDigital - *Digitalização do Diccionario da Lingua Portugueza de António de Morais Silva* [PTDC/LLT-LIN/6841/2020] is a project financed by the Portuguese National Funding agency through the FCT – Fundação para a Ciência e Tecnologia. It started in March 2021 and will be founded till next year.
- MORDigital aims to make a historic Portuguese-language dictionary, the Morais dictionary, available as a digital resource.
- 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>

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.



Inria



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

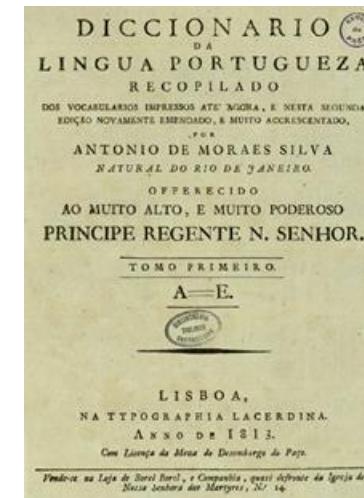
MORDigital project

The *Diccionario da Lingua Portugueza*

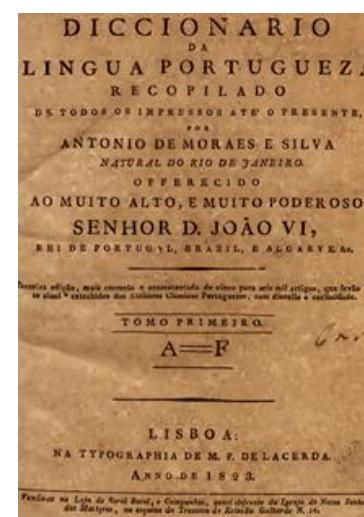
Frontispiece of Morais dictionaries (1789, 1813, 1823)



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.

Following best practices...

Establishing the data model

Modelling refers to how researchers conceptualise external representations – the process of creating a data model that can account for all the lexical data and their components.

Encoding refers to the process of expressing an abstract, conceptual model using a specific data format (e.g., TEI Lex-0).

- 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).
- For automatically structuring the OCRed dictionary pages into TEI Lex-0, GROBID-Dictionaries was chosen.

Refining metadata



- What is metadata?
 - data about data (literal meaning)
 - data that defines and describes other data (ISO/IEC 11179-1:2015)
 - record containing a description of a resource (ISO 24622-1:2015)
 - structured data set that describes and provides information about other data (ISO 1951 in revision)

Refining metadata: <teiHeader>

- 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.

```

</header>
<body>
<!-- Bilingual sections start here -->
<div type="section" n="1">
<p>fol taxado e feito Livro em papel a dous mil reis, Meia 8 de Junho de 1789.</p>
<p>em h[ab] remd "italic">Com tres rubricas.</h></p>
</div>
<!-- {...} -->
</body>
</TEI>

```

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

Refining metadata: abbreviations

```
<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">Braſt.</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">Filol.</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">Muſt.</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>

<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. uſt.</abbr>

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

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

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

Using consistent identifiers

- 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".

Enhancing the encoding

- 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 , tefouras , 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 papé-lão com repartimentos para navalhas , tefouras , facas , canivetes ,
&amp;
        </def>
    </sense>
    <pc></pc>
</entry>
```

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

Enhancing the encoding

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.

ABADERNAS, 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>ABADERNAS</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>
```

ABADERNAS

Enhancing the encoding

ABADA, s. f. A porção que leva a aba colhida, e apanhada § n. propr. 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. propr.</hi> de huma especie d'animal que tem ponta, e he o mesmo que <hi
rend="italic">Rinoceronte</hi></def>
        <pc>. </pc>
    </sense>
</entry>
```

ABADA.

Enhancing the encoding

ABCESSO. v. abſcesso.

```

<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">abſcesso</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



Enhancing the encoding

- TEI Lex-0 recommends that canonical labels should be defined in the `<teiHeader>` and then pointed to from the individual entries or senses in which these labels are used. Domain labels inside a sense are documented in `<encodingDesc>` (encoding description).

```
<encodingDesc>
  <classDecl>
    <taxonomy xml:id="domain">
      <category xml:id="domain.mathematical_sciences">
        <catDesc xml:lang="en">Mathematical Sciences</catDesc>
        <catDesc xml:lang="pt">Ciências Matemáticas</catDesc>
        <catDesc xml:lang="es">Ciencias Matemáticas</catDesc>
        <catDesc xml:lang="fr">Sciences mathématiques</catDesc>
      <category xml:id="domain.mathematical_sciences.mathematics">
        <catDesc xml:lang="en">Mathematics</catDesc>
        <catDesc xml:lang="pt">Matemática</catDesc>
        <catDesc xml:lang="es">Matemáticas</catDesc>
        <catDesc xml:lang="fr">Mathématiques</catDesc>
      <category xml:id="domain.mathematical_sciences.mathematics.arithmetic">
        <catDesc xml:lang="en">Arithmetic</catDesc>
        <catDesc xml:lang="pt">Aritmética</catDesc>
        <catDesc xml:lang="es">Aritmética</catDesc>
        <catDesc xml:lang="fr">Arithmétique</catDesc>
        <[...]>
      </category>
    </category>
  </taxonomy>
</classDecl>
</encodingDesc>
```

This hierarchical organisation constitutes the foundation of the domain ontology.

Enhancing the encoding

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">τ.</gram>
<gram type="gen">μ.</gram>
</gramGrp>
<sense xml:id="MORAIS.1.DLP.ABACO.s.1">
<!-- SEE usa-se na Architect. = domain --&gt;
&lt;def&gt;Peça superior do capitel da columna , serve como de
coberta ao cesto de flores , que nelle se representa ; usa-se
na&lt;/def&gt;
&lt;usg type="domain"&gt;Architect.&lt;/usg&gt;
&lt;/sense&gt;
&lt;metamark function="senseDelimiter"&gt;§&lt;/metamark&gt;
&lt;sense xml:id="MORAIS.1.DLP.ABACO.s.2"&gt;
&lt;usg type="domain"&gt;τ. arithm.&lt;/usg&gt;
&lt;def&gt;a taboada de Pythagoras&lt;/def&gt;
&lt;/sense&gt;
&lt;pc&gt;.&lt;/pc&gt;
&lt;/entry&gt;</pre>
```

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

Enhancing the encoding

ABAFADIÇO, adj. v. g. *lugar — calmofo*, em que não corre o ar livremente, ou viração *B. Pereira.* § *F. homem* — que se afronta facilmente. *Ulipo* 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 Ulipo -->
        <title>Ulipo</title>
        <citedRange unit="page">262</citedRange>
      </bibl>
    </cit>
  </sense>
</entry>
```

Enhancing the encoding

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. diasyntactic marking, namely domain labels, registers and part of speech categories).

Thank you!

... questions?

Ana Salgado: anasalgado@fcsh.unl.pt

Laurent Romary: laurent.romary@inria.fr

Rute Costa: rute.costa@fcsh.unl.pt

Toma Tasovac: ttasovac@humanistika.org

Anas Fahad Khan: fahad.khan@ilc.cnr.it

Margarida Ramos: mvramos@fcsh.unl.pt

Bruno Almeida: brunoalmeida@fcsh.unl.pt

Sara Carvalho: sara.carvalho@ua.pt

Mohamed Khemakhem: medkhemakhemfsegs@gmail.com

Raquel Silva: raq.silva@fcsh.unl.pt