



**HAL**  
open science

## The Morais Dictionary: Following Best Practices in a Retro-digitized Dictionary Project

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

► **To cite this version:**

Ana Salgado, Laurent Romary, Rute Costa, Toma Tasovac, Anas Fahad Khan, et al.. The Morais Dictionary: Following Best Practices in a Retro-digitized Dictionary Project. *International Journal of Humanities and Arts Computing*, 2024, 18 (1), pp.125 - 147. 10.3366/ijhac.2024.0325 . hal-04535611

**HAL Id: hal-04535611**

**<https://inria.hal.science/hal-04535611>**

Submitted on 6 Apr 2024

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












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



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

THE MORAIS DICTIONARY: FOLLOWING BEST PRACTICES IN A  
RETRO-DIGITIZED DICTIONARY PROJECT

---

ANA SALGADO , LAURENT ROMARY , RUTE COSTA ,  
TOMA TASOVAC , ANAS FAHAD KHAN ,  
MARGARIDA RAMOS , BRUNO ALMEIDA ,  
SARA CARVALHO , MOHAMED KHEMAKHEM ,  
RAQUEL SILVA  and BORIS LEHEČKA 

**Abstract** *This article outlines essential best practices for retro-digitized dictionary projects, using the ongoing MORDigital project (DOI 10.54499/PTDC/LLT-LIN/6841/2020) as a case study. The MORDigital project focuses on digitally transforming the historically significant Portuguese Morais dictionary's first three editions (1789, 1813, 1823). While the primary objective is to create faithful digital versions of these renowned dictionaries, MORDigital stands out by going beyond the mere adoption of established best practices. Instead, it reflects on the choices made throughout the process, providing insights into the decision-making process. The key topics emphasized include (1) the establishment of a robust data model; (2) the refinement of metadata; (3) the implementation of consistent identifiers; and (4) the enhancement of encoding techniques; additionally exploring the issue of structuring domain labelling. The article aims to contribute to the ongoing discourse on best practices in retro-digitized dictionary projects and their implications for data preservation and knowledge organization.*

---

*International Journal of Humanities and Arts Computing* 18.1 (2024): 125–147  
DOI: 10.3366/ijhac.2024.0325

© Ana Salgado, Laurent Romary, Rute Costa, Toma Tasovac, Anas Fahad Khan, Margarida Ramos, Bruno Almeida, Sara Carvalho, Mohamed Khemakhem, Raquel Silva and Boris Lehečka. The online version of this article is published as Open Access under the terms of the Creative Commons Attribution-NonCommercial Licence (<http://www.creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, distribution and reproduction provided the original work is cited. For commercial re-use, please refer to our website at: [www.eupublishing.com/customer-services/authors/permissions](http://www.eupublishing.com/customer-services/authors/permissions).  
[www.eupublishing.com/ijhac](http://www.eupublishing.com/ijhac)

**Keywords:** digital preservation, historical dictionary, knowledge organization, lexicography, retro-digitization, standards

## I. INTRODUCTION

The MORDigital<sup>1</sup> project, funded by the Portuguese Fundação para a Ciência e Tecnologia (FCT), is dedicated to creating digital versions of the first three editions<sup>2</sup> of a historically significant Portuguese dictionary—the Morais dictionary. While our primary goal is to make these editions accessible on the Web, MORDigital goes beyond mere digitization. We combine established best practices with innovative computational methods to create a user-friendly web interface powered by TEI Publisher.<sup>3</sup>

This research addresses a significant gap in Portuguese-speaking countries, where there is a lack of searchable online retro-digitized dictionaries that adhere to current data-sharing standards.<sup>4</sup> Furthermore, the pipeline used in MORDigital and the practical insights gained from working with historical dictionaries will benefit other researchers undertaking similar tasks.

Moreover, MORDigital operates in an era marked by the emergence of mass digitization projects, revolutionizing the field of lexicography. These initiatives aim to preserve heritage materials and simplify access to historical documents, emphasizing the importance of best practices in ensuring data quality and integrity. Technological progress has taken us beyond mere scanned texts. Our ultimate goal is to transform printed dictionaries into computational, readily exploitable lexical resources, enhancing searchability.<sup>5</sup>

This shift underscores the increasing significance of using advanced technologies and standards to enhance the accessibility of digitized dictionaries for research and academic purposes. While practical applications for retro-digitized historical dictionaries may not always be immediately evident, our primary mission is to elevate their value within the scholarly community. This includes tracing the evolution of language, identifying obsolete lexical units, exploring shifts in meaning and usage over time, and investigating the terminology used in specific historical periods to provide a deeper understanding of the past.

We recognize that ‘best practices’ are well-researched procedures, backed by evidence and experience, designed for broad adoption to ensure optimal working methods. MORDigital is just one piece of a broader landscape, where numerous historical dictionaries are undergoing digitization processes and becoming accessible online. Noteworthy projects and standards, such as Nénufar,<sup>6</sup> BASNUM<sup>7</sup> and eDIL,<sup>8</sup> among others, have played an instrumental role in shaping the project’s approach. These collaborative efforts are expected to significantly enrich research and foster a more comprehensive understanding of linguistic and cultural evolution across various historical periods.

Throughout this article we will explore the intricacies of the encoding process, shedding light on the significant decisions made along the way. In essence, this article serves as a roadmap, emphasizing the importance of best practices as thoughtful choices, rather than mere standards to be followed. We aim to achieve our goals of interoperability and long-term sustainability in retro-digitized dictionary projects by addressing pivotal topics such as (1) the establishment of a robust data model, involving the creation of a structured framework to systematically organize and represent data; (2) the refinement of metadata, enriching data with descriptive information that elucidates its context and augments searchability; (3) the implementation of consistent identifiers, a challenging task given their pivotal role in preserving data integrity, and particularly complex in the context of machine-readable dictionaries; and (4) the enhancement of encoding techniques, optimising data storage and information retrieval through efficient encoding methods.

## 2. THE MORAIS DICTIONARY

The *Dicionario da Lingua Portuguesa*, by António de Morais Silva, also known as the Morais dictionary, was first published in 1789. This lexicographic work is a pivotal milestone in Portuguese monolingual lexicography and effectively represents the dawn of a new era in that field.

As part of its remarkable legacy, this dictionary laid the foundation for subsequent Portuguese lexicographic work throughout the nineteenth and twentieth centuries. Developed during the Enlightenment period, the Morais dictionary served as a model for the ensuing modern general language dictionaries.

The first edition of the Morais dictionary comprises two volumes: volume 1 with 752 pages and volume 2 with 541 pages. It is worth noting that Morais does not claim authorship but attributes the work to Rafael Bluteau, acknowledged as the author of *Vocabulario Portuguez e Latino*. However, Morais explicitly acknowledges in the 'Prólogo ao Leitor' (Prologue to the Reader) that his contributions to the dictionary are important. The second (1813) and third (1823) editions, each published in two volumes, are considered distinct dictionaries due to their extensive enrichment and updates (Figure 1). It is essential to clarify that Morais assumes authorship in the second edition. The digitized versions of the dictionary, which are presently accessible as PDF files in the public domain,<sup>9</sup> underwent a rigorous re-OCRization process to guarantee that the project's ultimate output meets high-quality standards.

## 3. THE ESTABLISHMENT OF A ROBUST DATA MODEL

Adopting existing data models has several advantages for lexicographic datasets, especially for interoperability<sup>10</sup> with other existing systems and resources and for



Figure 1. Frontispieces of the Moraes dictionary (1789, 1813, 1823).

supporting the content's longer-term sustainability.<sup>11</sup> A well-defined data model ensures consistency in how the retro-digitized dictionary data is structured and organized. A well-defined structure establishes clear guidelines for capturing and representing lexicographic content, ensuring uniformity across different projects. Adhering to established standards and formats by following the FAIR<sup>12</sup> principles makes the data compatible or interoperable and can be combined or linked with other lexical or lexicographic datasets. It also enables efficient search and retrieval of lexicographic content, making it easier for end-users to navigate and explore the digital edition. We must add that a robust data model supports the long-term sustainability of retro-digitized dictionary projects.

Concerning retro-digitized dictionaries, two primary initiatives have gained significant levels of adoption: first, the TEI Guidelines and its dedicated module for dictionaries found in Chapter 9 ('Dictionaries').<sup>13</sup> Within this scope, the TEI Lex-0,<sup>14</sup> a customization of the original TEI guidelines supported by the DARIAH Working Group Lexical Resources, offers another approach specifically designed for machine-readable dictionaries; second, the Lexicon Model for Ontologies (Ontolex-Lemon), specifically the Lexicography Module (*lexicog*<sup>15</sup>), is embraced by the Ontolex-Lexicon Community group.

The research community and several organizations have widely taken up the TEI guidelines. TEI Lex-0, meanwhile, serves as both a technical specification and a set of community-driven recommendations for encoding lexicographic resources. While TEI primarily focuses on lexicographic resources as digital editions, Ontolex serves as the reference model for encoding these resources as Linked Data.

In the case of the Morais editions, the TEI Lex-0 encoding has been completed, and we will now proceed with the conversion to RDF based on Ontolex-Lemon.<sup>16</sup> This approach combines the advantages of both TEI and Ontolex, enabling comprehensive representation and interoperability of retro-digitized dictionaries, because TEI enables a thorough and structured mark-up of textual content, while Ontolex facilitates representation as Linked Data, allowing the content to be integrated with data from external lexical resources, which may not be limited to dictionaries.

#### 4. THE REFINEMENT OF METADATA

According to ISO/IEC 11179-1:2015,<sup>17</sup> metadata is defined as 'data that defines and describes other data'. ISO 24622-1:2015 provides another significant definition, referring to it as a 'record containing a description of a resource'. This definition highlights the essential role of metadata in capturing detailed information about a resource, further emphasizing the importance of refining metadata to enhance data description and management. Metadata can be

understood as a structured dataset that describes and provides information about other data.

When working with lexicographic resources, this refined definition encompasses the components unique to lexicographic resources, allowing for a more targeted and precise understanding of metadata within this context. The refinement enhances the descriptive and explanatory aspects of data, making the retrieval of relevant information more efficient and accurate compared to the previous state. This improved organization and categorization of data enable users to navigate and access specific pieces of information. Moreover, refining metadata promotes consistency in managing information across various resources. While it may be challenging to ensure, establishing standardized formats, vocabularies and metadata structures makes integrating and exchanging data across different platforms and domains easier compared to the absence of such structured data. This, in turn, facilitates improved interoperability.

Overall, refining metadata allows user-friendly information retrieval and fosters uniformity in managing data resources. To achieve this goal, it is essential to identify the relevant metadata elements that are appropriate for describing a certain resource. This involves using established schemas to ensure that the metadata is structured and organized in a consistent and standardized way.

To highlight the significance of metadata, we will specifically focus on the TEI header,<sup>18</sup> which is a central element within the structure of any TEI document. The TEI header serves as a structured container and repository for the metadata and additional information associated with the encoded text. Specifically, in the case of the Morais dictionary, the encoding process starts with the <teiHeader> element. The <teiHeader> element contains comprehensive bibliographic data on both the printed source(s) and the electronic file. The inclusion of detailed bibliographic information within the TEI header will facilitate end-users' interactions with the Morais digital versions.

Following the TEI Lex-0 specification (version: 0.9.2<sup>19</sup>), the TEI header of the Morais dictionary comprises three essential components:

- **file description** (<fileDesc>): This mandatory section provides a comprehensive bibliographic description of both the machine-readable resource and the original analogue source.
- **encoding description** (<encodingDesc>): This section includes crucial encoding principles and decisions made during the process. It also encompasses the taxonomy of domain labels, which describes the classification used within the dictionary.
- **profile description** (<profileDesc>): The profile description specifies the object and working languages employed in the dictionary. It is important to note that the first edition of the Morais dictionary consists of two volumes (A–K and L–Z).

As a result, each volume is represented by its dedicated <biblStruct> element within the *file description/source description*. This approach guarantees accurate differentiation and well-organized bibliographic data.

By adhering to the TEI Lex-0 specification, the Morais dictionary's TEI header incorporates these key components, ensuring a thorough and standardized representation of metadata for the resource.

The significance of documenting typographical conventions should also be highlighted here. Printed dictionaries, as physical books with finite dimensions, call for developing various strategies and conventions that now characterize them as textual resources. Typography has played a pivotal role in the dissemination of dictionaries, serving multiple purposes: (1) saving space (e.g., strategies such as using abbreviated forms have been employed to optimize space utilization within the limited confines of a printed dictionary); (2) enhancing structure and accessibility (e.g., the use of paragraph indentation and bold typefaces to highlight lemmas, or headwords, in dictionary articles facilitates easy identification and navigation for users); 3) providing labels (e.g., usage labels, such as register markers including colloquial or other restrictions associated with headwords, convey valuable information to users). These labels aid in understanding and contextualizing the usage of specific lexical units.

Abbreviations in printed dictionaries are typically compiled in a list located in the front matter. It is sometimes desirable to mark abbreviations in the copy text, whether to trigger special processing for them, to provide the full form of the unit abbreviated, or to allow for different possible expansions of the abbreviation. Abbreviations may be transcribed as they stand, using the <abbr> tag, or expanded, using the <expn><sup>20</sup> tag.

The @type attribute can further distinguish different types of abbreviations based on their intended function. Within the MORDigital project, we have identified and categorised six distinct values. These values encompass various aspects of linguistic information and offer valuable insights into the microstructure of the dictionary. The identified values are as follows:

- **POS (part-of-speech)**: This value indicates the grammatical category to which a particular lexical unit belongs, such as noun, verb or adjective;
- **usage** (domain; time; geographic; sociocultural; textType; frequency): These values encompass multiple dimensions, including domains, time, geographic location, sociocultural context, text type and frequency, providing insights into the usage patterns;
- **gender**: This value represents the grammatical gender of nouns, adjectives or pronouns;
- **number**: The number value denotes the grammatical number of nouns or pronouns, specifying whether they are singular or plural;



```

<abbr type="domain">Agric.</abbr>
<abbr type="domain">Anat.</abbr>
<abbr type="domain">Archit.</abbr>
<abbr type="domain">Arithm.</abbr>
<abbr type="domain">Articul.</abbr>
<abbr type="domain">Aftrol.</abbr>
<abbr type="domain">Afttron.</abbr>
<abbr type="domain">Botan.</abbr>
<abbr type="domain">Braf.</abbr>
<abbr type="domain">Chim.</abbr>
<abbr type="domain">Cirurg.</abbr>
<abbr type="domain">Chron.</abbr> | <abbr type="domain">Cron.</abbr>
<abbr type="domain">Defalt.</abbr>
<abbr type="domain">Filof.</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">Jurifp.</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">Muf.</abbr>
<abbr type="domain">Naut.</abbr>
<abbr type="domain">Opt.</abbr>
<abbr type="domain">Sortogr.</abbr>
<abbr type="domain">Perfp.</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">Aflat.</abbr>
<abbr type="time">Ant.</abbr> | <abbr type="time">antiq.</abbr>
<abbr type="text">Poet.</abbr>
<abbr type="socioCultural">Sch.</abbr> | <abbr type="socioCultural">Schul.</abbr>
<abbr type="socioCultural">Fan.</abbr>
<abbr type="socioCultural">Walg.</abbr>
<abbr type="frequency">Freq.</abbr>
<abbr type="frequency">P. uf.</abbr>
<abbr type="gender">Com.</abbr>
<abbr type="gender">F.</abbr>
<abbr type="number">P1.</abbr>
<abbr type="number">Sing.</abbr>

```

**Figure 2.** Example of the different values found.

- **categorization** (verbs subcategories; degrees of adjectives): This value encompasses subcategories related to verb forms and categories associated with degrees of adjectives, shedding light on the specific grammatical variations and nuances within the dictionary entries;
- **hint**: The hint value provides additional indications or references within the dictionary, such as chapter indications or other relevant contextual information found within the text.

By recognizing and categorising these six values, the MORDigital project aims to enrich the overall richness and usability of the dictionary, allowing users to explore and analyse the lexicographic data from diverse linguistic perspectives (see Figure 2).

In addition to the considerations mentioned earlier, gathering information about various typographical conventions employed in the dictionary is crucial. These conventions include different delimiters and their respective functions, which are vital in organizing and structuring the content. We suggest using the <metamark> element to encode and represent these delimiters accurately. Some examples of these delimiters are:

- "lemmaDelimiter": This delimiter is used after a lemma, marking the end of the lemma;
- "posDelimiter": Employed after the part of speech, the posDelimiter indicates the boundary between the part of speech and the subsequent information related to the article;
- "usageDelimiter": This is placed before a usage label;

- "senseDelimiter": This is positioned before a new sense, marking the transition between different senses or meanings of a lexical unit.

By documenting these typographical conventions, the MORDigital project ensures the preservation of the intricate details found in printed dictionaries and their accurate representation in digital formats. This meticulous documentation contributes to a faithful and comprehensive portrayal of the original source, maintaining the integrity of the dictionary's structure and enabling its effective utilisation in the digital environment.

## 5. THE IMPLEMENTATION OF CONSISTENT IDENTIFIERS

Establishing a consistent identification system for content is crucial for maximizing its reusability and simplifying effective referencing. To achieve this, it is recommended to define different levels of granularity. In our approach, we utilize the `xml:id` attribute, which requires a unique value within the XML document. For uniqueness and clarity, we use a dot (.) as a delimiter for subsequent identifier parts. Automated creation of these unique ids is accomplished by implementing an XSLT script. The generated id comprises several components, including the author's name, abbreviated dictionary title, edition number and a non-accented lemma (e.g. "MORAIS.DLP.1.ABA") and an accented lemma (e.g."MORAIS.DLP.1.BÁCORO"). This combination of elements results in a distinctive and meaningful identifier for each lexicographic article. In the case of homonyms, we add abbreviations of grammatical categories to distinguish between similar items separated by an underscore: "MORAIS.DLP.1.CELIBATO\_s-m" and "MORAIS.DLP.1.CELIBATO\_adj." Concerning hyphenated compounds, the hyphens are part of the id, too ("MORAIS.DLP.1.JACTAR-SE").

Through the consistent application of this identification system, we enhance the discoverability and interoperability of the encoded content, facilitating efficient cross-referencing and promoting the reuse of data within the digital environment. We have also implemented a similar system for other sections of the dictionary, such as the list of abbreviations. Using a very granular system of identifiers will likely ensure the uniqueness of identifiers across multiple dictionaries.

## 6. THE ENHANCEMENT OF ENCODING TECHNIQUES

To automatically structure the OCRed dictionary pages into TEI Lex-0 format, we have opted to use GROBID-Dictionaries.<sup>21</sup> In the planning phase of the MORDigital project, an important decision was made to preserve the textual content precisely as it appeared in the original printed edition. This commitment

```
<entry xml:id="MORAIS.DLP.1...." type="..." xml:lang="...">
  <form type="lemma">
    <orth>...</orth>
  </form>
  <metamark function="lemmaDelimiter">,</metamark>
  <gramGrp>
    <gram type="pos"/>
    <gram type="gen"/>
  </gramGrp>
  <sense xml:id="MORAIS.DLP.1....sense...">
    [...]
  </sense>
</entry>
```

**Figure 3.** Basic structure of a lexicographic article in the Morais Silva (1789) dictionary.

to preserving the exact textual content is invaluable for scholars, enabling them to study dictionary articles in their original context.

In the Morais dictionary, each lexicographic article, represented by an `<entry>` element, generally (there are some exceptions) starts with a lemma, also known as the canonical form. The lemma is encoded using the `<form>` element, which includes the `@type` attribute set to "lemma". The `<orth>` element captures the orthographic form of the lemma – that is, the written form per se. Subsequently, all grammar-related elements are encoded within the `<gramGrp>` element, while elements on meanings are enclosed within the `<sense>` element. In summary, the basic structure of a lexicographic article (as illustrated in Figure 3) looks like this:

The basic structure of encoding, exemplified in Figure 4, demonstrates a practical implementation using a real example extracted from the Morais dictionary:

The lexical unit ESTOJO [case; cover; kit] exhibits several traditional typographic characteristics in the Morais dictionary. The article begins with the headword, presented in uppercase letters, followed by a comma as a conventional typographic delimiter preceded by a space. The abbreviated grammatical information ‘f. m.’ [masculine noun] is then provided, and finally, the lexicographic definition ‘caixinha de coiro, ou papêlão com repartimentos para navalhas, tesouras, facas, canivetes, &c.’ [a small case or box made of leather or cardboard with compartments for razors, scissors, knives, penknives, etc.]. To represent the comma and its function within the dictionary, we encode

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

```
<entry xml:id="MORAIS.DLP.1.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.DLP.1.ESTOJO.sense.1">
    <def>caixinha de coiro , ou papélão com repartimentos para navalhas
, tefouras , facas , canivetes , &amp;c</def>
  </sense>
  <pc>.</pc>
</entry>
```

**Figure 4.** ESTOJO [case; cover; kit], an example of a basic article structure in the Morais Silva (1789) dictionary.

it as a `<metamark>` element. In this particular case, the `<metamark>` serves as a label, indicating the separation between the form-related and grammar-related elements. Alternatively, it can be considered as a `<lbl>` (label) denoting the comma as a delimiter or simply as a punctuation character (`<pc>`). This encoding approach allows for a precise representation of the typographic conventions and their respective functions within the lexicographic article.

Occasionally—it is important to note that we are working with a historical dictionary dating back to the eighteenth century—instead of the comma serving as the delimiter, a semicolon is used (Figure 5). It is essential to highlight that the TEI element encoded remains unchanged.

Another variation that may occur is the placement of the grammatical number before the gender (see Figure 6):

Multiple senses within a dictionary entry are typically distinguished using symbols or markers. In the Morais dictionary, only the second and following senses are delimited by the symbol §; the first one is not (see Figure 7):

With regard to cross-references, it is customary for them to be introduced by the abbreviation *v.* (which stands for *veja*, meaning ‘see’ in the context), as illustrated in Figure 8. This convention serves to direct the reader’s attention to related information or articles, simplifying navigation and access to additional relevant content:

**JALDE ; adj. còr amarella acceza.**

```
<entry xml:id="MORAIS.DLP.1.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.DLP.1.JALDE.sense.1">
    <def>còr amarella acceza</def>
  </sense>
  <pc>.</pc>
</entry>
```

**Figure 5.** JALDE [yellow colour], an example of an article with a semicolon delimiting the lemma from the POS in the Morais Silva (1789) dictionary.

**ABADERNÁS, plur. femin. naut. ganchos onde fe fixáo os colhedores, e outros cabos, quando fe aberta a enxarcia.**

```
<entry xml:id="MORAIS.DLP.1.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.DLP.1.ABADERNAS.sense.1">
    <usg type="domain">naut.</usg>
    <def>ganchos onde fe fixáo os colhedores, e outros cabos, quando fe aberta a enxarcia</def>
  </sense>
  <pc>.</pc>
</entry>
```

**Figure 6.** ABADERNAS [rope hook], an example of an article where the number appears first, followed by the gender in the Morais Silva (1789) dictionary.

Cross-references are more comprehensive in certain instances, including additional information such as synonymic definitions to address potential ambiguities. This approach aims to provide readers with a more comprehensive

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

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

Figure 7. ABADA [folded flap], an example of an article with two senses in the Morais Silva (1789) dictionary.

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>
```

Figure 8. ABCESSO [abscess], cross-reference preceded by a v. in the Morais Silva (1789) dictionary.

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

```

<entry xml:id="MORAIS.DLP.1.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.DLP.1.ABADEJO.sense.1">
    <xr type="related">
      <lbl expand="veja">
        <hi>v.</hi>
      </lbl>
      <ref target="#MORAIS.DLP.1.VACA-LOURA" type="mainEntry">Vaca
loura</ref>
    </xr>
  </sense>
  <metamark function="senseDelimiter">:</metamark>
  <sense xml:id="MORAIS.DLP.1.ABADEJO.sense.2">
    <xr type="related">
      <lbl expand="veja:">
        <hi>v.</hi>
      </lbl>
      <ref target="#MORAIS.DLP.1.ABADEJO" type="mainEntry">
        <hi>Badejo</hi>
      </ref>
    </xr>
  </sense>
</pc>.</pc>
</entry>

```

**Figure 9.** ABADEJO [stag beetle], cross-reference preceded by a *v.*, followed by a synonymic definition, a colon, *v.* and another cross-reference in the Morais Silva (1789) dictionary.

understanding of the referenced term or concept, improving clarity and minimizing potential confusion (Figure 9):

## 6.1. Encoding Domain Labels

When it comes to classifying dictionaries, it's important to note that classifications can vary among different authors.<sup>22</sup> The most traditional classification distinguishes two major categories: (1) language dictionaries and (2) encyclopaedic dictionaries, which combine linguistic and non-linguistic information. In our work, we have classified Morais dictionaries as general language dictionaries because they compile, preserve and describe the lexicon of the Portuguese language, as opposed to specialized language dictionaries,

which focus on specific aspects of linguistic description, such as a portion of the lexicon or a particular thematic area. Notably, these dictionaries are historical works from past centuries. The Morais dictionaries encompass a comprehensive collection of lexical units, ranging from general to specialized units. In our study, we specifically focus on these specialized lexical units or terms. Consequently, we place significant emphasis on diatechnical information, particularly domain labels.

These labels serve as markers that associate a specific lexical unit with a particular domain or field of knowledge. To achieve this, we adopt an onomasiological approach,<sup>23</sup> taking the concept and its respective concept system as the central elements of terminological work applied to general language dictionaries. Our proposal revolves around the organization and conceptualization of knowledge within general language dictionaries, using the Morais dictionary as a case study. Through this approach we recognize the benefits of implementing a hierarchical structure within the numerous labels employed, leading to enhanced information retrieval capabilities.

Concerning the organization of domains, we recognize the importance of hierarchical levels, namely the superdomain, domain and subdomain.<sup>24</sup> These levels provide a systematic framework for categorizing and structuring the diverse domains encountered in the dictionary. The superdomain represents the highest level of taxonomic grouping, encompassing broad and overarching categories. Below the superdomain we have the domain level, which delves deeper into specific subject areas or fields. Finally, the subdomain level further refines the classification, representing narrower subsets within a broader domain.

TEI Lex-0 provides valuable restrictions and guidelines for accurately encoding domain labels. In line with the TEI Guidelines and common practice across various projects, it is considered best practice to use the `<usg>` element with the `@type` attribute to designate it as a domain label. In TEI Lex-0, the `@type` attribute is compulsory and must adhere to predefined values. Additionally, TEI Lex-0 offers a variety of sample values of the `@type` attribute to demonstrate the potential applications of the typed element `<usg>`. Notably, TEI Lex-0 introduces a revised naming scheme for the sample values<sup>25</sup> in the original TEI Guidelines. For instance, it replaces the former abbreviation "dom" with the more explicit "domain" to enhance clarity, as seen in the updated tag: `<usg type="domain"/>`.

The domain label can be represented as an abbreviated form, a very commonly used lexicographic convention for usage information in dictionary systems due to space restrictions in print dictionaries or expanded. The Morais dictionary uses abbreviated forms whose expansion is provided in the initial pages of the dictionary. When encoding dictionary data, it is important to normalize the abbreviated and unabbreviated labels to a single value for the sake of consistency and better information retrieval, as already mentioned.



A good practice is to encode the abbreviated domain label within the element `<usg>`, followed by the attribute type required (`@type="domain"`).

While our primary objective is not to achieve immediate overall harmonization, the development of a multilingual domain map has prompted us to introduce additional metadata to enhance our analysis. As part of this process, we assigned metalabels to the respective domains, indicating their equivalent English terms. This approach enhances our ability to examine and compare domains across different languages, facilitating a more comprehensive data analysis.<sup>26</sup>

To address the limitation of a flat representation of labels in general language dictionaries – the initial list is static and just in alphabetic order – we aim to adopt an encoding approach in which we can separate canonical, possibly multilingual, labels that are defined in one place and then simply pointed to from the dictionary entry. Ideally, we aim for a system where these labels, including multilingual ones, can be centrally defined and referenced throughout the dictionary. To achieve this, we suggest using the current mechanism for defining taxonomies within the `<teiHeader>`. This approach will allow us to create a structured framework for organizing and linking labels, thereby improving the overall organization and accessibility of the dictionary's content.

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). The `<taxonomy>` element identifies the structured taxonomy. The categories are documented in the `<category>` element. The category elements are described, each defining a single category within the given taxonomy. Then, child categories are defined by the contents of a nested `<catDesc>` (category description) element, which contains the designation of the domain in question in the identified language. A single category may contain more than one `<catDesc>` child; accordingly, the categories can be described in different languages (`@xml:lang`). The `<term>` (term) contains the designation which identifies the field of special knowledge (domain). The `<gloss>` (gloss) identifies a definition, a description or an explanation of the domain. As a result of this thought process, we can establish a multilingual hierarchy for the MATHEMATICAL SCIENCES superdomain (Figure 10):

This hierarchical organization constitutes the foundation of the domain ontology of MATHEMATICAL SCIENCES,<sup>27</sup> also developed within the scope of the MORDigital project.

We will now demonstrate the encoding process in TEI Lex-0 for the lexicographic article ABACO (Figure 11) within the Morais dictionary. As previously elucidated, TEI Lex-0 incorporates encoding usage information by implementing the `<usg>` typed element. Specifically, in the case of domain

```

<encodingDesc>
<!-- [...] -->
<!-- Hierarchical domain labels -->
  <taxonomy xml:id="domains">
    <category xml:id="domain.mathematical_sciences"
      valueDatcat="http://www.semanticweb.org/OntoDomLab-Math#MathematicalSciences
      http://vocabs.rossio.fcsh.unl.pt/moreis_domains/0036">
      <catDesc xml:lang="en">
        <term>Mathematical Sciences</term>
        <gloss>Group of areas of study that includes, in addition to mathematics,
        those academic disciplines that are primarily mathematical in nature but may not
        be universally considered subfields of mathematics proper.</gloss>
      </catDesc>
      <catDesc xml:lang="pt">
        <term>Ciências Matemáticas</term>
      </catDesc>
      <category xml:id="domain.mathematics"
        valueDatcat="http://www.semanticweb.org/OntoDomLab-Math#Mathematics
        http://vocabs.rossio.fcsh.unl.pt/moreis_domains/0024">
        <catDesc xml:lang="en">
          <term>Mathematics</term>
        </catDesc>
        <catDesc xml:lang="pt">
          <term>Matemática</term>
        </catDesc>
        <category xml:id="domain.arithmetic"
          valueDatcat="http://www.semanticweb.org/OntoDomLab-Math#Arithmetic
          http://vocabs.rossio.fcsh.unl.pt/moreis_domains/0003">
          <catDesc xml:lang="en">
            <term>Arithmetic</term>
          </catDesc>
          <catDesc xml:lang="pt">
            <term>Aritmética</term>
          </catDesc>
        </category>
        <category xml:id="domain.geometry"
          valueDatcat="http://www.semanticweb.org/OntoDomLab-Math#Geometry
          http://vocabs.rossio.fcsh.unl.pt/moreis_domains/0018">
          <catDesc xml:lang="en">
            <term>Geometry</term>
          </catDesc>
          <catDesc xml:lang="pt">
            <term>Geometria</term>
          </catDesc>
        </category>
      </category>
    </taxonomy>
  </classDecl>
</encodingDesc>

```

Figure 10. The hierarchical domain labels for MATHEMATICAL SCIENCES.

ABACO, f. m. Peça superior do capitel da columna, ferve como de cobertura ao cefto 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="MOR1.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="MOR1.DLP.ABACO.s.1">
    <!-- SEE usa-se na Architect. = domain -->
    <def>Peça superior do capitel da columna , ferve como de cobertura ao cefto
de flores , que nelle se representa</def>
    <metamark function="domainDelimiter">;</metamark>
    <usg type="domain">usa-se na <hi rend="italics">Architect.</hi></usg>
  </sense>
  <metamark function="senseDelimiter">§</metamark>
  <sense xml:id="MOR1.DLP.ABACO.s.2">
    <usg type="domain" rend="italics">t. arithm.</usg>
    <def>a taboada de Pythagoras</def>
  </sense>
  <pc>.</pc>
</entry>
```

Figure 11. ABACO [abacus], an example of a domain-labelled article.

labels, the value "domain" is assigned to the <usg> element. Consequently, the explicit labels denoting the domains, as found within the article, are subsequently included as the content of the <usg> element. For instance, in the provided example, the labels *Architect.* and *t. arithm.* are part of the content of the <usg> element. In the first sense, the domain label appears after the lexicographic definition; in the second, the domain label is positioned before the corresponding meaning. The abbreviated classifier 't.' or 'Termo de...' [term of] is a textual marker referring to the domain in which a lexical unit was common in historical dictionaries.

In the given example, it is observed that the labels are interconnected by means of the @corresp attribute, establishing a linkage to their respective domains within the TEI-encoded taxonomy. These domains are specifically identified as "#domain.architecture" and "#domain.arithmetic". Additionally, the association of these domains with external Knowledge Organization Systems

(KOS) and ontologies is facilitated within the TEI header, wherein the taxonomy of domains is encoded.

Figure 11 exemplifies a segment of the taxonomy that pertains to the domains of architecture and arithmetic. Upon thoroughly examining the terminological work conducted thus far, it has been established that arithmetic is positioned hierarchically within the superdomain of MATHEMATICAL SCIENCES. On the other hand, the structuring of architecture is yet to be finalized, as it awaits the conclusive outcomes of subsequent terminological analysis.

To establish alignments between the taxonomy domains and classes within the OntoDomLab-Math ontology, explicitly focusing on MATHEMATICS and its subdomains (accessible at <https://github.com/moraisdigital/OntoDomLab-Math>), the @valueDatcat attribute is employed. Furthermore, the MorDigital Domain Classification concepts, formulated in SKOS, serve as reference points. For instance, the concept denoted by the URI [http://vocabs.rossio.fcsh.unl.pt/morais\\_domains/0003](http://vocabs.rossio.fcsh.unl.pt/morais_domains/0003) corresponds to the arithmetic domain.

This approach facilitates the TEI encoding of dictionary articles, allowing for alignment with external KOS and ontologies. Adopting this approach makes the encoding process more straightforward, thus improving the efficiency and effectiveness of representing dictionary articles within the TEI framework.

## 7. CONCLUSION












In conclusion, within the scope of the MORDigital project, we have identified a set of procedures to provide guidelines that contain the essential steps for achieving best practices. These guidelines aim to ensure the consistency and usability of the resulting lexicographic product. For data curation, it is imperative to: establish a robust data model, as well as a consistent framework for the organization of lexicographic content; use consistent terminology; refine metadata; and employ identifiers to support further the reliability and interoperability of the project. By enhancing encoding techniques, we can ensure an accurate representation of the lexical data.

The MORDigital project aims to enhance the presence of lexicographic digital content in Portuguese by collaborating with a broad community and adhering to standardized practices to facilitate linking these dictionaries with future resources. By applying a rigorous linguistic approach, the project enables the systematic organization and structural refinement of lexicographic components while also uncovering valuable lexical relationships among various elements. It's worth considering that the uniqueness of each ancient dictionary can make this a challenging task. However, we can still aim to establish linking mechanisms between the resulting structured dictionary and other resources, providing a prototype that may be adapted for future projects, particularly within the Portuguese-speaking community.

This project applies ontologies (e.g. diasystematic marking, namely domain labels, registers and part of speech categories) to facilitate a comprehensive analysis across the three editions of *Morais*. This approach enables a deeper understanding of the semantic and conceptual aspects of the dictionary's content. Using a metalabel will be beneficial for any work on aligning multiple dictionaries and studying them in parallel. However, an international harmonization effort across different dictionaries would require further comparison of more dictionaries and a community-based agreement on the common values for metalabels. The outcomes of this research contribute to the advancement of lexicography and pave the way for broader applications within the Portuguese language context, promoting further linguistic exploration and knowledge dissemination.

While our commitment involves fostering a transdisciplinary approach that integrates the principles and methodologies of terminology, lexicography and various other relevant disciplines, it is suggested that this approach could serve as an additional contribution to other current projects. We have made specific choices, but it is acknowledged that others may opt for equally valid alternatives. Embracing a holistic approach enables us to harness the strengths of each field and explore fresh perspectives and solutions. Our ultimate goal remains to support the broader objective of ensuring the ongoing accessibility and usability of retro-digitized dictionaries. We aspire to contribute to similar research projects by sharing our best practices and practical insights.

#### ORCID

Ana Salgado  <https://orcid.org/0000-0002-6670-3564>  
Laurent Romary  <https://orcid.org/0000-0002-0756-0508>  
Rute Costa  <https://orcid.org/0000-0002-3452-7228>  
Toma Tasovac  <https://orcid.org/0000-0002-3919-993X>  
Anas Fahad Khan  <https://orcid.org/0000-0002-1551-7438>  
Margarida Ramos  <https://orcid.org/0000-0001-7209-3806>  
Bruno Almeida  <https://orcid.org/0000-0002-5777-5574>  
Sara Carvalho  <https://orcid.org/0000-0002-7501-5405>  
Mohamed Khemakhem  <https://orcid.org/0000-0003-3529-2990>  
Raquel Silva  <https://orcid.org/0000-0002-0505-4863>  
Boris Lehečka  <https://orcid.org/0000-0003-4893-5537>

#### ACKNOWLEDGEMENTS

This work is supported by (1) the MORDigital–Digitalização do *Diccionario da Lingua Portuguesa* de António de Morais Silva (DOI 10.54499/PTDC/LLT-LIN/6841/2020) project financed by the Portuguese National Funding through the FCT–Fundação para a Ciência e Tecnologia; (2) 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.

END NOTES

- <sup>1</sup> *MORDigital*, <https://mordigital.fcsh.unl.pt/>, last accessed 5 November 2023.
- <sup>2</sup> A. Morais Silva, *Diccionario da lingua portugueza composto pelo padre D. Rafael Bluteau, reformado, e accrescentado por Antonio de Moraes Silva, natural do Rio de Janeiro*, vols 1–2 (Lisbon, 1789); A. Morais Silva, *Diccionario da lingua portugueza recopilado dos vocabulários impressos até agora, e nesta segunda edição novamente emendado, e muito accrescentado* (Lisbon, 1813); A. Morais Silva, *Diccionario da lingua portugueza recopilado de todos os impressos até o presente* (Lisbon, 1823).
- <sup>3</sup> *TEI Publisher*, <https://teipublisher.com/>, last accessed 5 November 2023.
- <sup>4</sup> R. Costa, A. Salgado, F. Kahn, S. Carvalho, L. Romary, B. Almeida, M. Khemakhem, M. Ramos, R. Silva and T. Tasovac, ‘MORDigital: the advent of a new lexicographical Portuguese project’, *Electronic lexicography in the 21st century: post-editing lexicography, Proceedings of the eLex 2021 Conference* (Brno, 2021), 321–4.
- <sup>5</sup> S. R. Perry, ‘Digitization and digital preservation: a review of the literature’, *SLIS Student Research Journal*, 4, no. 1 (2014), <http://scholarworks.sjsu.edu/slissrj/vol4/iss1/4>, last accessed 5 November 2023; G. Budin, S. Majewski, and K. Mörth, ‘Creating lexical resources in TEI P5’, *Journal of the Text Encoding Initiative*, 3 (2012), doi:10.4000/jtei.522, last accessed 15 November 2023; M. Rundell, ‘From print to digital: implications for dictionary policy and lexicographic conventions’, *Lexikos*, 25, no. 1 (2015), 301–22, doi:10.5788/25-1-1301, last accessed 15 November 2023; C. Tiberius, J. Kallas, S. Koeva, M. Langemets and I. Kosem, ‘A lexicographic practice map of Europe’, *International Journal of Lexicography* (2023), ecad023, doi:10.1093/ijl/ecad023, last accessed 15 November 2023.
- <sup>6</sup> H. Bohbot, F. Frontini, G. Luxardo, M. Khemakhem and L. Romary, ‘Presenting the Nénufar Project: a diachronic digital edition of the Petit Larousse Illustré’, *GLOBALEX 2018–Globalex workshop at LREC2018* (Miyazaki, 2018), 1–6, <https://hal.science/hal-01728328/>. The website of Nénufar (Nouvelle édition numérique de fac-similés de référence) is available here: <http://nenufar.huma-num.fr/presentation/>, last accessed 5 November 2023.
- <sup>7</sup> Agence nationale de la recherche, *Funded projects*, <https://anr.fr/Project-ANR-18-CE38-0003>, last accessed 5 November 2023; G. C. Williams, I. Galleron and C. Stincone, ‘Announcing the dictionary: front matter in the three editions of Furetière’s Dictionnaire universel’, *Proceedings of the XIX EURALEX Congress* (Greece, 2020), 393–402; Clarissa Stincone, ‘Usage labels in Basnage’s Dictionnaire universel (1701)’, in A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs and P. Storjohann, eds, *Proceedings of the XX EURALEX International Congress* (Mannheim, 2022), 755–64, <https://shs.hal.science/halshs-04016166>, last accessed 5 November 2023.
- <sup>8</sup> eDIL 2019: *An Electronic Dictionary of the Irish Language*, based on the Contributions to a Dictionary of the Irish Language (Dublin, 1913–76), [www.dil.ie](http://www.dil.ie), last accessed 5 November 2023.
- <sup>9</sup> The PDF files of the two volumes of Morais’s first edition are available in the Portuguese National Library at the following link: <https://bndigital.bnportugal.gov.pt/records/item/79346-diccionario-da-lingua-portugueza-composto-pelo-padre-d-rafael-bluteau>. The second edition is here: <https://bndigital.bnportugal.gov.pt/records/item/258107-diccionario-da-lingua-portugueza?offset=5>, and the third one is here: <https://bndigital.bnportugal.gov.pt/records/item/258107-diccionario-da-lingua-portugueza?offset=5>, last accessed 5 November 2023.
- <sup>10</sup> According to ISO/IEC 2382:2015, ‘Information technology – vocabulary’, interoperability is defined as follows: ‘capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units’ (3).
- <sup>11</sup> R. Costa, C. Roche and A. Salgado, *Standards for representing lexicographic data: an overview. Version 1.0.0*, DARIAH-Campus, [Training module], (2022), <https://elexis.humanistika.org/id/REhOyKBu7pPs5zOAENDah>, last accessed 5 November 2023.

- <sup>12</sup> Findable, Accessible, Interoperable, Reusable; see M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak et al., ‘The FAIR guiding principles for scientific data management and stewardship’, *Scientific Data*, 3, 160018 (2016), doi:10.1038/sdata.2016.18, last accessed 5 November 2023.
- <sup>13</sup> TEI – Text Encoding Initiative, *P5: Guidelines for electronic text encoding and interchange*, <https://tei-c.org/release/doc/tei-p5-doc/en/html/DL.html>, last accessed 5 November 2023.
- <sup>14</sup> T. Tasovac, L. Romary et al., *TEI Lex-0: A baseline encoding for lexicographic data*. Version 0.9.1. DARIAH Working Group on Lexical Resources, (2018), <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>, last accessed 5 November 2023.
- <sup>15</sup> J. Bosque-Gil and J. Gracia (eds), *The OntoLex Lemon Lexicography Module* <https://www.w3.org/2019/09/lexicog/>, last accessed 5 November 2023; J. Bosque-Gil, J. Gracia and E. Montiel-Ponsoda, ‘Towards a module for lexicography in OntoLex’, *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets at 1st Language Data and Knowledge conference (LDK 2017)*, CEUR-WS, vol. 1899 (Galway, 2017), 74–84, [https://ceur-ws.org/Vol-1899/OntoLex\\_2017\\_paper\\_5.pdf](https://ceur-ws.org/Vol-1899/OntoLex_2017_paper_5.pdf), last accessed 5 November 2023.
- <sup>16</sup> B. Almeida, R. Costa, A. Salgado, M. Ramos, L. Romary, F. Khan, S. Carvalho, M. Khemakhem, R. Silva and T. Tasovac, ‘Modelling usage information in a legacy dictionary: from TEI Lex-0 to Ontolex-Lemon’, *COMHUM 2022 – 2nd Workshop on Computational Methods in the Humanities* (Lausanne, 2022), <https://run.unl.pt/handle/10362/151561>, last accessed 5 November 2023.
- <sup>17</sup> ISO/IEC 11179-1 (2015), ‘Information technology — Metadata registries (MDR) — Part 1: Framework. Geneva: International Organization for Standardization’.
- <sup>18</sup> GitHub, Morais Silva (1789) dictionary TEI header: [https://github.com/moraisdigital/MORDigital\\_teamwork/blob/main/TEI%20Lex-0%20encoding/header/TEI\\_Header\\_morais1\\_v7.xml](https://github.com/moraisdigital/MORDigital_teamwork/blob/main/TEI%20Lex-0%20encoding/header/TEI_Header_morais1_v7.xml)
- <sup>19</sup> For more detailed information about TEI Lex-0 constraints and recommendations related to the TEI header, see: <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#header>, last accessed 5 November 2023.
- <sup>20</sup> ‘The expan element may be used simply to record that an abbreviation has been silently expanded by the encoder.’ See <https://tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#CONAAB>, last accessed 5 November 2023.
- <sup>21</sup> M. Khemakhem, I. Galleron, G. Williams, L. Romary and P. J. O. Suárez, ‘How OCR performance can impact on the automatic extraction of dictionary content structures’, *19th Annual Conference and Members’ Meeting of the Text Encoding Initiative Consortium* (Graz, 2019), <https://hal.science/hal-02263276/document>, last accessed 5 November 2023.
- <sup>22</sup> A. Rey, ‘Typologie génétique des dictionnaires’, *Langages*, 19 (1970), 48–68; L. Zgusta, *Manual of lexicography* (Prague and The Hague, 1971); B. Svensén, *Practical lexicography: principles and methods of dictionary-making* (Oxford, 1993); S. I. Landau, *Dictionaries: the art and craft of lexicography* (Cambridge, 2001); B. T. S. Atkins and M. Rundell, *The Oxford guide to practical lexicography* (New York, 2018).
- <sup>23</sup> R. Costa, A. Salgado, M. Ramos, B. Almeida, R. Silva, S. Carvalho, F. Khan, T. Tasovac, M. Khemakhem and L. Romary, ‘A crossroad between lexicography and terminology work: knowledge organisation and domain labelling’, *Digital Scholarship in the Humanities*, 38 (Supplement 1) (2023), i17–i29, doi:10.1093/llc/fqad022, last accessed 5 November 2023.
- <sup>24</sup> A. Salgado, ‘Terminological methods in lexicography: conceptualising, organising and encoding terms in general language dictionaries’ (PhD thesis, Universidade Nova de Lisboa, Faculdade de Ciências Sociais e Humanas, 2021), <https://run.unl.pt/handle/10362/137023>, last accessed 5 November 2023.

- <sup>25</sup> A. Salgado, R. Costa and T. Tasovac (2019) 'Improving the consistency of usage labelling in dictionaries with TEI Lex-0', *Lexicography*, 6 (2) (2019), 133–56, doi10.1007/s40607-019-00061-x, last accessed 5 November 2023.
- <sup>26</sup> See GitHub, *MORDigital*: <https://github.com/moraisdigital/MORDigital/tree/main/Domain%20Labelling>, last accessed 5 November 2023.
- <sup>27</sup> R. Costa, A. Salgado, F. Kahn, M. Ramos, S. Carvalho, T. Tasovac, B. Almeida, M. Khemakhem, L. Romary and R. Silva, 'Integrating terminological and ontological principles into a lexicographic resource', *International Conference, Multilingual digital terminology today. Design, representation formats and management systems* (Padua, 2022).