HUGOD Conference
Humanities Going Digital: Teaching, training & research experiences

# Morais dictionary: following best practices in a retrodigitised dictionary project

Salgado, Ana[1,2], Romary, Laurent[3], Costa, Rute[1], Tasovac, Toma[4], Kahn, Fahad[5], Ramos, Margarida[1], Almeida, Bruno[1], Carvalho, Sara[1,6], Khemakhem, Mohamed[7], Silva, Raquel[1]

[1] CLUNL – Centro de Linguística da Universidade NOVA de Lisboa
[2] Academia das Ciências de Lisboa
[3] ALMAnaCH – Automatic Language Modelling and ANAlysis & Computational Humanities Inria de Paris
[4] BCDH – Belgrade Center for Digital Humanities
[5] CNR-ILC – Istituto di Linguistica Computazionale "Antonio Zampolli"
[6] CLLC – Centro de Línguas, Literaturas e Culturas
[7] ArcaScience

**Keywords:** digital preservation; historical dictionary; retrodigitisation; standards

This paper aims to describe some of the best practices that should be followed in any retrodigitised dictionary project, taking as an example an ongoing project – MORDigital[1]. We will focus our attention on the importance of 1) establishing the data model; 2) refining metadata; 3) using consistent identifiers; 4) enhancing the encoding.

The MORDigital project, funded by the Portuguese Fundação para a Ciência e Tecnologia, concerns the analysis and enrichment of a legacy Portuguese dictionary, namely the first three editions of the Morais dictionary, in order to test innovative computational digital methods and to make these editions available online in a browsable web interface. In the Portuguese context, this research fills a gap concerning searchable online retrodigitised dictionaries, built on current standards and methodologies which promote data sharing and harmonisation (Costa et al., 2021). Moreover, the pipeline used in the MORDigital, as well as our more general practical observations of working with historical dictionaries, should be useful for anyone working on similar tasks (Khan et al., 2022).

Currently, a great number of historical dictionaries are being digitised and made available online (e.g. MORDigital, Nénufar[2], BASNUM[3], eDIL[4], CDSL[5], among many others), which represents an excellent opportunity to compare their structure and content with the ultimate goal of linking these different lexicographic resources.

[1] https://mordigital.fcsh.unl.pt/
[2] http://nenufar.huma-num.fr/presentation/
[3] https://anr.fr/Project-ANR-18-CE38-0003
[4] https://dil.ie/
[5] https://www.sanskrit-lexicon.uni-koeln.de/scans/csldev/csldoc/build/index.html

For this paper, after a brief introduction concerning the need to preserve cultural heritage and its sustainable management, we will cover the following topics:

**1) Establishing the data model.** For lexicographic datasets, adopting existing data models has several advantages, especially for interoperability reasons and also for supporting the longer-term sustainability of the content (Costa et al., 2022a). Concerning retrodigitised dictionaries, there are two initiatives which, in general, have been adopted: 1) the TEI Guidelines and its specific module for dictionaries in Chapter 9 ('Dictionaries')[6] by TEI Consortium or TEI Lex-0 customisation hosted by the DARIAH Working Group Lexical Resources and 2) the Lexicon Model for Ontologies (Ontolex-Lemon), namely the Lexicography Module (lexicog)[7] from the Ontolex-Lexicon Community group. The TEI Guidelines have been widely used by research communities and some organisations. TEI Lex-0, meanwhile, is both a technical specification and a set of community-based recommendations for encoding machine-readable dictionaries. While TEI represents the dictionary as a digital edition, Ontolex is the reference model for the encoding of dictionaries as linked open data. The encoding of the Morais' editions will be carried out in TEI Lex-0 and then converted to RDF based on Ontolex-Lemon.

**2) Refining metadata.** ISO/IEC 11179-1 (2015) defines metadata as 'data that defines and describes other data' (p. 3). Refining metadata enables information retrieval and brings consistency to the management of all types of information.

As an example of metadata, we will point to the TEI header, a key element of the structure of any TEI document which contains additional information about the encoded text. The Morais dictionary encoding starts with the `<teiHeader>` element (Fig. 1), in which the metadata 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.

[6] https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html
[7] https://www.w3.org/2019/09/lexicog/

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

Applying the TEI Lex-0 specification[8], the Morais dictionary TEI header consists of 1) a `<fileDesc>`, which is mandatory and presents the complete bibliographic description of the machine-readable resource and of the analog original source; 2) an `<encodingDesc>`, where we add information about some important principles and decisions taken during the encoding and where the taxonomy of domain labels is described; and 3) a `<profileDesc>`, where we specify the object and working languages.

Typographical conventions should also be highlighted. The fact that a printed dictionary is a book with finite dimensions led to developing several strategies and certain conventions that characterise it as a text today. Undoubtedly, the typographic technique served multiple purposes: (1) to save space (e.g., space-saving devices such as abbreviated forms); (2) to reflect and facilitate access structure (e.g., bold typefaces to signal the lemma in a dictionary article are easier to find); (3) labels to inform the user about certain restrictions of the headword (e.g., usage labels). Abbreviations may be transcribed as they stand, `<abbr>`, or expanded, `<expan>`. The `<type>` attribute may be used to distinguish types of abbreviations by their function. In the MORDigital,

[8] 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

we identified 7 different values: POS; domain; usage; gender; number; grammar; miscellaneous.

```
<abbr type="POS" norm="adjective">adj.</abbr>
<abbr type="domain">Agric.</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="frequency">Freq.</abbr>
<abbr type="gender">Com.</abbr>
<abbr type="number">Pl.</abbr>
<abbr type="grammar">At.</abbr>
<abbr type="hint">C.</abbr> || <abbr type="hint">Cap.</abbr>
```

Fig. 2: Example of the different values found

It is also important to collect other conventions such as the different delimiters and their function throughout the dictionary, e.g., `"lemmaDelimiter"`, `"posDelimiter"`, `"usageDelimiter"` and `"senseDelimiter"`.

**3) 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"`.

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

One of the first decisions taken in the planning of the MORDigital project was to keep the textual content exactly as it appeared in the printed edition. Fig. 3 shows an example of the textual content of a basic structure of a lexicographic article:

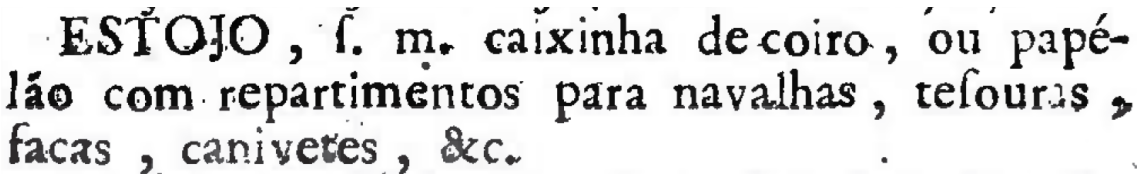
ESTOJO, ſ. m. caixinha de coiro, ou papelão com repartimentos para navalhas, teſouras, facas, canivetes, &c.

Fig. 3: ESTOJO [case; cover; kit], an example of a basic article structure.

We will explain in detail some of the important choices made during the encoding. Furthermore, we will further explore, throughout this article, the application of terminological methods to lexicographic work by combining semasiological/onomasiological approaches, thereby providing added value via the use of ontologies (Costa et al., 2022b).

In sum, we intend to provide a general overview of some of the pertinent topics addressed within the scope of this project in order to achieve the desired interoperability and the longer-term sustainability of digitised content.

## References

Costa et al. (2022a). Standards for Representing Lexicographic Data: An Overview. Version 1.0.0. DARIAH-Campus. [Training module].

Costa et al. (2022b). Integrating terminological and ontological principles into a lexicographic resource. International Conference, Multilingual digital terminology today. Design, representation formats and management systems, Università degli Studi di Padova.

Costa et al. (2021). MORDigital: the advent of a new lexicographical Portuguese project. Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference., Lexical Computing CZ s.r.o., Brno, Czech Republic, pp. 321–324. ISSN 2533-5626.

ISO/IEC 11179-1 (2015). Information technology — Metadata registries (MDR) — Part 1: Framework. Geneva: International Organization for Standardization.

Khan et al. (2022). Interlinking lexicographic data in the MORDigital project. LLODREAM2022: LLOD approaches for language data research and management. Mykolas Romeris University, Vilnius, Lithuania.