

A crossroad between lexicography and terminology work: knowledge organisation and domain labelling

Rute Costa, Ana Salgado, Margarida Ramos, Bruno Almeida, Raquel Silva
CLUNL – Centro de Linguística da Universidade NOVA de Lisboa

Sara Carvalho
CLUNL – Centro de Linguística da Universidade NOVA de Lisboa
CLLC – Centro de Línguas, Literaturas Modernas

Fahad Khan
CNR-ILC – Istituto di Linguistica Computazionale “Antonio Zampolli”

Toma Tasovac
BCDH – Belgrade Center for Digital Humanities

Mohamed Khemakhem
ArcaScience

Laurent Romary
ALMA^{na}CH – Automatic Language Modelling and ANALysis & Computational Humanities Inria de Paris

Abstract

MORDigital project aims to encode the selected editions of *Diccionario de Lingua Portuguesa* by António de Morais Silva, first published in 1789. Our ultimate goals are, on the one hand, to promote accessibility to cultural heritage while fostering reusability and, on the other hand, to contribute towards a more significant presence of lexicographic digital content in Portuguese through open tools and standards. The Morais dictionary represents a significant legacy, since it marks the beginning of Portuguese dictionaries, having served as a model for all subsequent lexicographic production. The team follows a new paradigm in lexicography, which results from the convergence between lexicography, terminology, computational linguistics, and ontologies as an integral part of digital humanities and linked (open) data. 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, namely TEI Lex-0. The team will further ensure the connection to other existing systems and lexical resources, particularly in the Portuguese-speaking world.

Résumé

Le projet MORDigital vise à encoder les éditions du Dicionario de Lingua Portuguesa d'António de Morais Silva, publié pour la première fois en 1789. Les objectifs ultimes sont, d'une part, de promouvoir l'accessibilité au patrimoine culturel tout en favorisant la réutilisation et, d'autre part, contribuer à une présence plus significative du contenu numérique lexicographique en portugais à travers des outils en libre accès et des standards. Le dictionnaire Morais représente un patrimoine important, puisqu'il marque le début des dictionnaires portugais, ayant servi de modèle à toute la production lexicographique ultérieure. L'équipe suit un nouveau paradigme en lexicographie, qui résulte de la convergence entre la lexicographie, la terminologie, la linguistique computationnelle et les ontologies en tant que partie intégrante des humanités numériques et des données (ouvertes) liées. Dans le contexte portugais, cette recherche comble une lacune concernant les dictionnaires rétronumérisés consultables en ligne, construits sur des normes et des méthodologies actuelles qui favorisent le partage et l'harmonisation des données, à savoir TEI Lex-0. L'équipe assurera en outre la connexion aux autres systèmes et ressources lexicales existants, en particulier dans le monde lusophone.

1. Introduction

MORDigital¹ aims to supply high-quality digital versions of successive editions (1789, 1813, 1823) of the *Dicionario de Lingua Portuguesa*, by António de Morais Silva (Morais Silva, 1789), a heritage object, which will be converted into structured data. The project aims to ensure their interoperability with other existing systems and resources by converting them into structured data via TEI Lex-0² (a simplified sub-format of TEI, serialised in XML, for encoding dictionaries), LMF (Romary, Khemakhem, Khan et al., 2019) (an ISO standard for the integration of a wide variety of electronic lexical resources) and Ontolex-Lemon³ (a *de facto* standard created in 2016 to represent lexical information in RDF format). The value of these standards is evident by the role they play in interoperability at the semantic level, by allowing the integration of the resource's data and metadata in the Linguistic Linked Open Data Cloud (LLOD).⁴

From the very beginning, the TEI Guidelines have had a module explicitly focused on the encoding of dictionaries. However, this module has been criticised for its extreme flexibility, i.e., the existence of multiple possibilities to encode similar structures that affect the interoperability of the encoded formats. In some cases, TEI makes no binding requirements for the possible values since there are many possibilities across different

projects. To reduce this freedom and define a specific format for dictionaries, forcing dictionary encoders to follow the same structural rules, the lexicographic and dictionary-encoding communities are currently discussing TEI Lex-0 (Tasovac, Romary, Bánski et al., 2018) with a particular focus on retro-digitised dictionaries, and, in this paper, we will deal exclusively with this new format. Complying with TEI Lex-0 specifications allows us to find solutions to cover all the microstructural elements of the dictionary. We found some advantages in the application of TEI Lex-0 that we sum up as follows: 1) it represents an excellent opportunity to define a metalanguage suitable for the encoding of lexicographic components; 2) the accuracy of the encoding, reducing possible cases of ambiguity; 3) significant constraints are crucial, as some TEI practices can compromise the desired interoperability; 4) the verbosity favours more detailed and linguistically appropriate encoding.

We intend to apply methodologies regarding the computer-assisted reading of the text with a TEI-compliant format, using advanced techniques to turn the Morais dictionary into a computer-readable resource.⁵ The output of the work will be made available via a dedicated platform whose construction is in progress.

Interoperability requires a prior linguistic analysis of the metalinguistic classification of the data comprising the microstructure of the various editions. In the context of this paper, we analysed the flat domain labels listed in the front matter of

¹ <https://mordigital.fcsh.unl.pt/en/about/>

² <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

³ <https://www.w3.org/2019/09/lexicog/>

⁴ <https://linguistic-lod.org/>

⁵ The project has been extensively described in:

https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_18_pp312-324.pdf

the Morais dictionary. We concentrated our efforts on the terms accompanied by domain labels belonging to domains that can be grouped under the umbrella concept MATHEMATICAL SCIENCES. The purpose of the article is to propose a methodology to reorganise the set of identified and related domain labels in a hierarchy. This reorganisation is formally defined and is a necessary step for the ontologisation of knowledge, allowing, in the end, the alignment of the label-related metadata in the editions of the Morais dictionary.

This paper is organised as follows: the first (and current) section introduces and outlines the article. Section 2 reviews the theoretical framework. In Section 3, we present the methodology that has been used to deal with the domain labels. Section 4 introduces the Morais dictionary. Section 5 describes the OCRisation process, as the pre-phase for applying GROBID-Dictionaries⁶. Section 6 introduces the domain labelling issue as a device to identify specialised lexical content in general language dictionaries. We end the paper by highlighting the importance of combining lexicographic work with terminological methodologies, namely linking the dictionary content to ontologies, in Section 7.

2. Theoretical Framework

Lexicography and Terminology are two related disciplines, sharing interests in the study of terms and various types of meanings (definitions, examples, among others) related to such terms, with the ultimate goal of helping understand, conceptualise and share knowledge. Both disciplines have as their final aim to supply lexical resources in which the lexical data are organised and structured for specific end-users and purposes. Albeit sharing the same objects of study – terms and related information –, both disciplines differ mainly because they apply different methodologies, grounded on distinct frameworks as they meet specific social demands, being resources designed and produced to respond to diverse needs.

While lexicographic work is concerned with designing and producing dictionaries to keep track of lexical units in use, thereby describing them in all their facets, terminology work is ‘concerned with the systematic collection, description, processing and presentation of concepts and their designations’

(ISO 1087:2019, p. 13) in a domain. Systematicity and structured knowledge are key elements in terminology work, since a term is a designation of a concept that belongs to a concept system. Considering that general language dictionaries contain terms, we argue that terminological principles must be applied to lexicographic work.

The specificity of a terminological dictionary is to be a structured collection of terminological articles, where the headword always holds a term serving as an entry point into a terminological article. In this case, the terminologist wants to give an answer to the questions ‘what is x ’, ‘what is the function of x ’ or ‘what is the composition of x ’, for instance. On the other hand, a language dictionary is a ‘lexicographic resource containing a structured collection of lexicographic articles’ (Costa, Roche and Salgado, 2022), where the headword always holds a lexical unit, which can be a term or not.

Whilst the lexicographic methodology follows a semasiological path, in the sense that it departs from an existing corpus of lexical units to explore their semantic values, the terminological methods first try to identify the concepts and subsequently order the terms found by reference to a concept system, thereby following an onomasiological approach and resorting to the construction of conceptual representations of the domains under analysis. These different approaches should not be seen as polar opposites; in fact, quite the contrary: ‘la perspective linguistique, plutôt sémasiologique et la perspective conceptuelle, plutôt onomasiologique, [...] ne s’excluent pas mutuellement, mais se complètent’ [the linguistic perspective, which is more semasiological, and the conceptual perspective, which is more onomasiological, [...] are not mutually exclusive; they are complementary] (Costa, 2006, p. 85).

Morais is a language dictionary. This means that the microstructure is composed of a structured set of lexical units, both general and specialised. We have therefore decided to focus on a portion of lemmas that are terminological units (terms). Taking this into consideration, we resort to an onomasiological approach, taking the concept and its respective concept system as the central elements of terminological work applied to general dictionaries. In dictionaries, labels are markers that indicate a restricted use of a lemma, whereas domain labels – the labels that we are working on in this article – are ‘markers that identify the specialised field of

⁶ <https://github.com/MedKhem/grobid-dictionaries>

knowledge in which a lexical unit is mainly used' (Salgado, Costa and Tasovac, 2019) and are therefore considered to be terms.

Our approach assumes that terminological and lexicographic approaches can be complementary. Lexicography, terminology – when dealing with born-digital, retrodigitised or historical resources –, ontologies and computational linguistics can be considered integral parts of the digital humanities, necessarily implying a paradigm shift in the construction of dictionary resources. Hence, the increasing relevance of following compatible standards and formats when working on lexical data.

3. Research Methodology

The three steps described in this paper constitute the preparatory stage for the remaining parts of the methodology. In the first step of the procedure, we used OCR to digitise the dictionary sources (cf. section 5). Given that the high-quality OCRisation is proceeding at a good pace, we started the second step of the method with the lexicographic analysis of the macro and microstructure of the dictionary. In the macrostructure, we analysed the set of lemmas and proceeded with the survey of the components of the microstructure, where we identified and selected the lemmas marked as domains belonging to the MATHEMATICAL SCIENCES (cf. section 6). After selecting the flat list of these labels, we built a hierarchical structure to organise the domain (cf. section 6.3), and moved forward with the hierarchical domain labels encoding process using the TEI Lex-0 specification. The third step of the methodology corresponds to the building of the domain ontology (cf. section 7).

The three steps described in this paper are the preparatory stage for moving on to the remaining parts of the methodology, which are the automatic structuring of the lexical content for the creation of a computer-readable resource. The Morais's digitised versions will be structured using GROBID-Dictionaries, an open-source machine learning system for the parsing, extraction and structuring of lexical information obtained from dictionary text.

Afterwards, we will convert and map the TEI content to the LMF standard and their respective serialisations, as well as to OntoLex-Lemon. The final aim is to link the data and align the lemmas, senses and other lexicographic content between the

three editions that will be available on a platform for Morais, enriched with both lexicographic and ontological modules.

4. Morais Dictionary

The first edition of the *Diccionario da Lingua Portuguesa* in 1789 (Morais Silva, 1789), authored by António de Morais Silva, commonly known as the Morais dictionary, marks the beginning of the modern, monolingual contemporary Portuguese lexicography. This dictionary represents a significant legacy, having served as a model for all subsequent lexicographic Portuguese production throughout the 19th and 20th centuries. This dictionary was devised during the Enlightenment and was influenced by other modern language dictionaries published in Europe in the 16th and 17th centuries.

The first edition of the dictionary presents two volumes (vol. 1, 752 pp. and vol. 2, 541 pp.). Morais does not claim to be the author, assigning this condition to Bluteau, 'the author of the Vocabulário Portuguez and Latino'. Morais recognises, however, in the 'Prólogo ao Leitor' [Prologue to the Reader] that the additions he brought to the dictionary are quite relevant. The second (1813) and third (1823) editions are considered new dictionaries, due to both their enrichment and updating. Nonetheless, it should be made clear that in the second edition, Morais already assumes himself as the author. The digitised versions of the dictionary, available in the public domain as PDF files, are currently undergoing a re-OCRisation process to ensure the quality of the final output of the project.

5. OCRisation Process

Retrodigitising historical dictionaries into machine-readable dictionaries poses several challenges that the scientific community has tried to resolve through the creation of tools, different formats and the establishment of standards for modelling lexical resources and making them available. Our starting point was the set of digitised files available as PDFs in the public domain. As high-quality digitisation is required in order to use GROBID-Dictionaries (Khemakhem, Foppiano, Romary, 2017; Khemakhem, Galleron, Williams et al., 2019), we decided to re-OCRise the files.

The process of re-OCRisation to ensure the quality of the text, i.e., without noise such as ink stains,

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; i	instead of long /s/	f		
ll; jj; (T;	instead of double long /s/	ff		
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 / ' /		ä	A79 (10); A80 (3); A81 (13); A76 (6); A77 (8); A78 (5); A85 (4); A91 (4); A92 (6); A93 (8); A94 (11)
<i>Adadeira</i>	the 1st letter of most capitalised forms in italic is replaced by 2 letters, e.g., / <i>Ad</i> / instead of / <i>M</i> /		<i>Madeira</i>	
<i>Fleira</i>	the 1st letter of a specific proper noun in italic is recursively replaced, i.e., / F/ instead of / V /		<i>Vleira</i>	
d; if; ff; ff;	ligature e.g. [activamente]: activamente.		ct	
r	typographic		t	

Table 1. Some common recurrent types of noise caused by the optical character recognition tool

missing characters – given the age of the original printed document –, or misrecognition of old characters by the OCR tool⁷ – just to name a few – implies several manual activities: (1) cleaning the common errors generated by the tool, (2) replacing the misrecognised characters, (3) inserting missing characters/text and finally (4) printing in PDF the page where the cleaning tasks occurred. During these activities, we have observed that the OCR tool creates recurrent types of noise, e.g., the old character /l/ [long /s/] is generally replaced by /f/ or frequently replaced by /i/; linguistic forms are updated to their contemporary spelling form if a graphic accent is currently used (e.g., *alguem* vs. *alguém*).

The existence of recurrent types of noise as shown by the examples in Table 1 allows us to use the

common feature find > replace a given character/word – a strategy that partially accelerates the cleaning tasks. The average time to conclude a page is approximately 30 minutes if no other issues arise after printing the outcome of the cleaning tasks. We are referring to the text structure that sometimes gets distorted by the OCRisation (e.g., the indentation suffers misalignment due to an ink spot), an issue that affects the outcome saved in PDF format – a format that is parameterised to print by default as an ‘exact copy’ of the original document. In our case, the exact copy mirrors the text structured in 2 columns. Along with the exact-copy-PDFs, we further decided to save the outcomes in ‘formatted text’ – a format that structures the text in 1 column, hence losing the original text structure. This option allows us to manually validate the previous format, such as the absence of hyphen translineation. Hyphenation is a critical issue in this project, given the large number of hyphenated words in Portuguese (e.g., reflexive verbs and their clitics), in addition to the 2-column layout. The causes of silent characters are varied, yet the age of the document mainly leads to the erasure of smaller characters. Some of these issues are illustrated in Fig. 1, where the lexicographic article of PASTOR [shepherd] is shown before and after the re-OCRisation. For the case of a silent hyphen, see the reflexive verb ‘defendê lo’ [defend him]; it should be ‘defendê-lo’.

To surmount issues related to unrecognised words, or words hidden by ink spots, among other unclarities, we resort to the later editions of Morais, namely the 2nd edition from 1813, or the 3rd, from 1831.

The outcomes of the re-OCRisation, namely the PDFs (two text formats of the same page) and corresponding FineReader files (the actual work in the OCR tool) are being uploaded in a collaborative environment cloud. In that same environment, the

Original text	Re-OCRisation before cleaning the OCR noise	Print after cleaning the OCR noise
<p>PASTOR, f. m. o que guarda, e apalcanta o gado. § f. <i>Pajlor</i>, o Cura d'almas, e todo o ministro da Igreja, que adminiltra o palto espi-ritual. § O Rei como diz Homero deve fer pajlor do feu povo, i. e. adminiltrar-lhe, de que viva farto defendê lo dos inimigos internos, e ex-ternos; e tirar delle fô o que bafiar para as neccellidades suas, e do público. <i>Barros. Elo-gio 1.</i></p>	<p>PASTOR, ri m. o que guarda, e apalcanta o gado. § ri <i>Pajlor</i>, o Cura d'almas, e todo o Wlinilfro da Igreja, que adminiltra o palto espi-ritual. § O Rei como tfa, Homero deve fer pajlor do feu povo, i. e. adtrimi ltrar-lhe, de que viva farto defendclo dos inimigos internos., e ex-ternos; e tirar delle fô o que bafiar para as neccellidadesi, suas, e do público. <i>Barras. -Elo-gio 1.</i></p>	<p>PASTOR, f. m. o que guarda, e apalcanta o gado. § f. <i>Pajlor</i>, o Cura d'almas, e todo o ministro da Igreja, que adminiltra o palto espi-ritual. § O Rei como diz Homero deve fer pajlor do feu povo, i. e. adminiltrar-lhe de que viva farto defendê lo dos inimigos internos, e ex-ternos; e tirar delle fô o que bafiar para as neccellidades suas, e do público. <i>Barros. Elo-gio 1.</i></p>

Fig. 1 Lexicographic article of PASTOR before and after cleaning the OCR noise

⁷ The OCR (optical character recognition) tool ABBYY (<https://www.abbyy.com/>).

team members are registering, on a spreadsheet, the observations (e.g., misspelling of a given word), along with the decisions (error updated or maintained) that can be useful in the next stage of the workflow (see next paragraph), plus the status of the work progression. The option of logging such observations resulted in the development of a useful state of affairs resource, given that the details of all pages and corresponding outcomes are indexed via hyperlinks – a feature that allows us to retrieve information efficiently.

The Morais’s digitised versions will then be structured using GROBID-Dictionaries. GROBID-Dictionaries takes as input lexical resources digitised in PDF format and generates a TEI-encoded hierarchy of the different recognised text structures. This software is used to parse the constituent parts of each lexicographic article, which involves the preparation of a native encoding format compliant with the XML/TEI metamodel. The model is very flexible and the result is a model of a historical dictionary whose entries are structured in a standard format, namely TEI Lex-0. We plan to adapt the system’s cascading architecture to allow the extraction of the different TEI constructs corresponding to the lexicographic structures and conventions. The outcome is a chain of cascading machine learning models, trained and evaluated against manually annotated data. This task will be started in the next phase of the project.

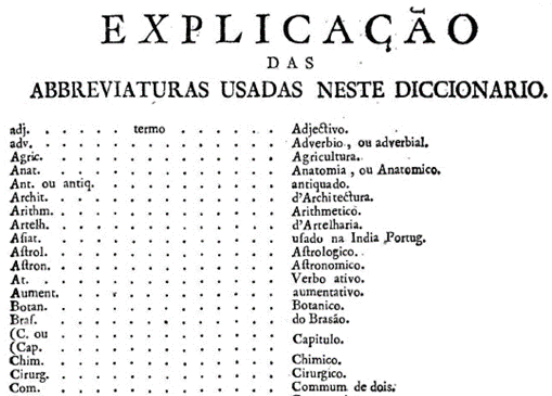
6. Domain Labelling

Usage labelling, a particular dictionary feature, plays a key role throughout this project. This type of labelling implies that ‘a certain lexical item deviates in a certain respect from the main bulk of items described in a dictionary’ (Svensén, 2009, p. 315).

We decided to focus on diatechnical information, which indicates that a given unit belongs to a particular domain. In the universe of the labelling system commonly used in lexicography, the labels assigned to these specialised senses are called domain labels. We use the term domain label not only to indicate abbreviations (e.g., Mathem.) collected in our dictionary corpus but also to refer to the extensions of each of the abbreviations written in full, e.g. ‘Mathematico’ [Mathematical]). As we are working on an 18th-century dictionary, the domain label designation is not the term related to the domain itself, but the adjectival form preceded by t. [abbreviation of term]. The classifier

termo is used as a textual marker, referring to the domain in which a lexical unit is used. Thus, ‘t. Mathematico’ must be understood as a term belonging to MATHEMATICS.

In the Morais dictionary, abbreviations are listed alphabetically in a section entitled ‘Explicação das abreviaturas usadas neste dicionario’ [Explanation of abbreviations used in this dictionary]. As seen in Fig. 2, there are two different columns: one containing the abbreviations and the other the domain designations written in full. The abbreviations are alphabetically ordered, without any particular concern about the relationships that can be established between the different types of labels.



E X P L I C A Ç Ã O	
D A S	
ABBREVIATURAS USADAS NESTE DICCIONARIO.	
adj.	termo
adv.	Adjectivo.
Agríc.	Adverbio, ou adverbial.
Anat.	Agricultura.
Ant. ou antiq.	Anatomia, ou Anatomico.
Archit.	antiquado.
Arithm.	d'Architectura.
Arithm.	Arithmetico.
Arithm.	d'Arithmetica.
Astrol.	ufado na India Portug.
Astrol.	Astrológico.
Astrol.	Astronomico.
At.	Verbo ativo.
Aument.	augmentativo.
Botan.	Botanico.
Bref.	do Brasão.
(C. ou	Capitulo.
(Cap.	Chimico.
Chim.	Cirurgico.
Cirurg.	Comum de dois:
Com.	

Fig. 2 Sample of the flat front matter list of Morais dictionary (Morais Silva, 1789, p. xi)

The analysis of the list of abbreviations allowed us to put forward a typology of usage information in the Morais dictionary: diatechnical, diatextual, diastatic, among others (Almeida, Costa, Salgado et al. 2022). In the next section, we will focus on diatechnical marking. Bearing in mind that knowledge is complex, Sager (1990) states, ‘In practice, no individual or group of individuals possesses the whole structure of a community’s knowledge; conventionally, we divide knowledge up into subject areas, or disciplines, which is equivalent to defining subspaces of the knowledge space’ (p. 16).

6.1 Domain Organisation

Domain organisation is crucial to improve the labelling system in dictionaries. Following an onomasiological approach, we propose to organise and conceptualise knowledge in general language dictionaries via the Morais dictionary case study. We see advantages of establishing a hierarchical

organisation within the multiple labels, thus improving information retrieval. This allows us to organise an increasing amount of terminological data, and to provide greater control over specialised content. As Silva (2014) states, ‘quanto melhor estiver organizado um sistema conceptual, mais fácil se torna, também, a gestão da terminologia’ (the better a concept system is organised, the easier it is to manage terminology; p. 135).

We noticed that some generic domains and subdomains coexist, including, for example, MATEMÁTICA [MATHEMATICS] and its subdomains ARITHMETICA [ARITHMETIC] and GEOMETRIA [GEOMETRY] or, for instance, MEDICINA [MEDICINE], together with CIRURGIA [SURGERY] and PHARMACIA [PHARMACY]. In this paper, we decided to explore MATHEMATICS and related mathematical domains.

In the organisation of domains, we consider the existence of three possible levels: superdomain, domain and subdomain (Salgado, 2021). The superdomain corresponds to the broadest taxonomic grouping, followed by the domain, whereas the subdomain is part of a broader domain.

6.2 Mathematical Sciences: A Case-Study

In the Morais dictionary, MATEMÁTICA [MATHEMATICS] is defined as ‘A sciencia, que ensina a conhecer as grandezas de toda sorte, suas razões, relações, e proporções: Mathematica mista (oppõe-se ás puras) a que ensina a aplicar os principios de calculo, e geometria aos corpos’ [The science, which teaches all kinds of quantities, their ratios, relations, and proportions: mixed Mathematics (as opposed to pure mathematics), which teaches how to apply the principles of calculus and geometry to bodies] (Morais Silva, 1789, vol. 2, p. 64). ARITHMETICA [ARITHMETIC] is presented as ‘Arte de calcular por algarismos’ [Art of calculating by numerals] (t. 1, p. 112), while GEOMETRIA [GEOMETRY] the ‘Parte da Mathematica, que ensina a conhecer a grandeza, razões, e proporções das grandezas continuas, ou sejam linhas, ou figuras, ou sólidos, ou superficies.’ [Part of Mathematics, which teaches the quantity, ratios, and proportions of continuous quantities, whether they are lines, or figures, or solids, or surfaces.] (Morais Silva, 1789, vol. 2, p. 86).

We selected three lexicographic articles to clarify and underly the rationale for the domain subdivisions, in which lemmas are mathematical

terms, namely terms used in GEOMETRY. We selected the following lexicographic articles: ACUTANGULO [acuteangle], DIAMETRO [diameter] and SO’LIDO [solid]. The application of a labelling system is not always entirely consistent in every dictionary, and even less so in historical dictionaries. Thus, as expected, we found a heterogeneous treatment for the selected terms.

ACUTANGULO, adj. Geometr. que tem tres angulos agudos v. g. *triangulo*.—

Fig. 3 Lexicographic article of ACUTANGULO (Morais Silva, 1789, vol. 1, p. 24)

In Fig. 3, the label Geometr. is used to mark the term acuteangle. Searching for diameter (Fig. 4), we also expected to find the domain label GEOMETRY.

DIAMETRO, f. m. a linha recta que tirada de hum ponto do circulo a outro passa polo seu ponto central. *P. Pereira* 2. f. 21. ufa deste termo significando a recta em contraposição da linha curva.

Fig. 4 Lexicographic article of DIAMETRO (Morais Silva, 1789, vol. 1, p. 435)

However, this lexicographic article is not marked.

Differently, the geometrical term <solid> (Fig. 5) – which also could be marked with the domain label GEOMETRY –, is marked with the MATHEMATICS domain label [Mathem].

SO’LIDO, adj. que não he fluido; o corpo cujas partes tem firme união, e não se de-tunem de si mesmas v. g. o pão, pedra, os metaes, &c. § Não fragil, que resiste ao em-bate, ou força sem se quebrar v. g. „ *sólido edificio, ponte sólida. Ulissea.* § f. Real, effectivo, duravel, que tem força, he bem fundado v. g. „ *doutrina*—; *amizade*—; *razões*—; *devoção*— § *Sólido*, em Mathem; se diz sub-tantivamente, o corpo que tem as 3 dimensões de largura, altura, e longor; oppõe-se a linha, e superficie. § *Número sólido*, v. cubico. § *Em sólido* v. *fóldum.* *F. Mendes* c. 151.

Fig. 5 Lexicographic article of SO’LIDO (Morais Silva, 1789, vol. 1, pp. 414–415)

These examples reveal the inconsistency and the lack of systematicity that may be encountered in general language dictionaries and even more so when working on a dictionary dating from the 18th century.

Assuming that the unlabelled lexical units belong to the general lexicon is controversial. In fact, not every lexical unit that can be considered a term is unlabelled. It is unclear if this is due to forgetfulness or if the lexicographer decided to apply different criteria.

To solve these issues, we are in favour of establishing a hierarchical organisation for the

multiple labels. To demonstrate our point of view, we proceed to the analysis of a concrete domain: what we consider MATHEMATICAL SCIENCES and all domains related to MATHEMATICS. To this end, we have consulted and analysed various classification systems to help us organise the domain labels related to MATHEMATICAL SCIENCES, arranging them according to what we consider to be the most appropriate dictionary labelling, given the absence of an explanation in the front matter on how the labelling system was applied. We decided to compare how other existing domain labelling classification systems organise their descriptors to establish analogies, particularly: the *Encyclopédie* of Diderot and d’Alembert, Dewey Decimal Classification (DDC), the UNESCO Thesaurus, EuroVoc and BabelNet.

We started by analysing the classification system used in the 18th century (when Morais was first published), namely the figurative system of human knowledge or the ‘Tree of Diderot and d’Alembert’, produced for the *Encyclopédie*. The three main branches of knowledge in the tree are: MÉMOIRE [MEMORY], HISTOIRE [HISTORY], RAISON [REASON], PHILOSOPHIE [PHILOSOPHY] and, finally, IMAGINATION/POÉSIE [IMAGINATION/POETRY].

According to the figurative system of human knowledge or the Tree of Diderot and d’Alembert, produced for the *Encyclopédie* by Jean le Rond d’Alembert and Denis Diderot⁸, the domain MATHÉMATIQUES [MATHEMATICS] is located under the hierarchy of RAISON [REASON]/SCIENCE DE LA NATURE [NATURAL SCIENCE], from where it divides

into three different categories: PURES [PURE], MIXTES [MIXED] and PHYSICOMATHÉMATIQUES [PHYSICOMATHEMATICS]. For this study, we are interested in locating ARITHMETIC and GEOMETRY. Both disciplines can be found under MATHEMATICS/PURE.

We continued to compare how other existing domain labelling classification systems organise their descriptors to establish analogies regarding these domains. In this work, we also considered: the Dewey Decimal Classification (DDC), the UNESCO Thesaurus, EuroVoc and BabelNet. The different classification proposals present hierarchical models ranging between domains and subdomains. After looking into the different classification systems, we chose to systematise the domains under study to find out their location and respective organisation. The outcome of this analysis is systematised in Table 2.

The first point to highlight is the similarity of the label treatment in all classification systems. In the Dewey Decimal Classification (DDC), MATHEMATICS is included in class 500, which is devoted to the broader class of NATURAL SCIENCE & MATHEMATICS. Specifically, MATHEMATICS is found in class 510, and is a kind of catchall for all the related sciences: ARITHMETIC (class 513) and GEOMETRY (class 516). Concerning EUROVOC, MATHEMATICS is found in the broader descriptor BT1 pure mathematics/BT2 mathematics. BabelNet considers that mathematics is a ‘science major universal’ and ARITHMETIC and GEOMETRY ‘pure mathematics area of mathematics’.

Morais	METALABEL	Encyclopedia of Diderot and d’Alembert	Dewey Decimal Classification (DDC)	EuroVoc	BabelNet
Arithm. Arithmetico	arithmetic	Reason Philosophy Science of Nature Mathematics Pure Arithmetics	500 Natural sciences & mathematics 510 Mathematics 513 Arithmetic	BT1 pure mathematics BT2 mathematics BT3 natural sciences	ISA pure mathematics area of mathematics
Geometr. Geometrico	geometry	Reason Philosophy Science of Nature Mathematics Pure Geometry	500 Natural sciences & mathematics 510 Mathematics 516 Geometry	BT1 pure mathematics BT2 mathematics BT3 natural sciences	ISA pure mathematics area of mathematics mathematics
Mathem. Mathematico	mathematics	Reason Philosophy Science of Nature Mathematics	500 Natural sciences & mathematics 510 Mathematics	BT1 natural sciences NT1 applied mathematics NT1 pure mathematics	ISA Science major Universal language Academic discipline formal science

Table 2 Comparison of Morais domain labels and existing classification systems

⁸https://www.wikiwand.com/en/Encyclop%C3%A9die#Media/File:ENC_SYSTEME_FIGURE.jpeg

Another point that we must pay attention to is the nature of the lexicographic works. In this case, we are dealing with general language dictionaries, not with terminological dictionaries. In principle, a greater degree of specialisation of a domain might require more knowledge of the end-user, but also more interpretation skills. However, some degree of specialisation might help the user better understand an entry within a given domain, when it is applicable. The subsequent definition should entail that information and be comprehensible and understandable for the end user. The organisation and subsequent segmentation of a domain as vast as that of MATHEMATICAL SCIENCES, in general, or MATHEMATICS, in particular, brings advantages for the end-user.

In Table 2, we have a column entitled metalabel, a tag that identifies the equivalent English designation of the corresponding domain. Using a metalabel will be beneficial for any work on aligning multiple dictionaries and studying them in parallel. This metalabel will also play an important role in the domain hierarchy that we will propose later for the benefit of annotation.

After comparing the different classification systems, we present our proposal to represent domains related to MATHEMATICAL SCIENCES (Fig. 6). We also use some anchors such as `hasSuperDomain`, `hasDomain` or `hasSubdomains`, which establish the relations between the different concepts.

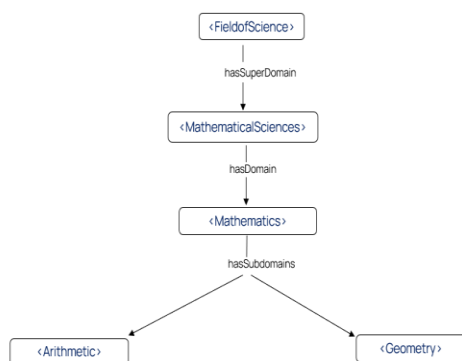


Fig. 6 Domain labels within the <MathematicalSciences> superdomain showing <Mathematics> as a domain and identifying its subdomains, <Arithmetic> and <Geometry>

In the proposal (Fig. 6), `<MathematicalSciences>` represents a broad subject area or a superdomain that can be decomposed into (hasDomain) a narrower subject branch (`<Mathematics>`). In turn, the narrower subject branch `<Mathematics>`

hasSubdomains: `<Arithmetic>` and `<Geometry>`. This example illustrates a generic-specific type of knowledge organisation, thereby allowing a transition from a non-hierarchical domain organisation to a hierarchical structure, which consequently increases the consistency of annotation and information retrieval. The hierarchical domain trees can be made visible to give end-users the possibility of understanding the conceptual scope and how terms are interlinked, since they are generally found isolated in general language dictionaries because they usually follow alphabetical order.

Even though MATHEMATICS as a domain label in general language dictionaries is part of a certain lexicographic tradition, we argue that MATHEMATICAL SCIENCES should be placed at the top level. Finally, it is self-evident that only the elaboration of concept systems will allow us to have a more concrete notion of the subdomains that should be conveyed and identify the many various concepts shared among the multiple subdomains.

The annotation of the superdomain, the domain and the subdomains will be made using TEI Lex-0 and will be explained in the next section.

6.3 Hierarchical Encoding of Domain Labels

The encoding of the previously established hierarchical domain labels is the following stage. Within the usage labels, as referred to above, the domain label is a crucial marker to identify terms in general language dictionaries. The restrictions that the TEI Lex-0 imposes on the TEI Guidelines are highly advantageous, as they allow a more precise and scientifically accurate encoding. It is considered good practice to restrict the scope of `<usg>`. The attribute `@type` must specify the element, in this case as a domain label. We decided to create new metadata, namely a metalabel, a tag that identifies the English equivalent of the corresponding domain. Using a metalabel will be beneficial for any work on aligning multiple dictionaries and studying them in parallel. However, an international harmonisation effort across different dictionaries would necessarily require further comparison of more dictionaries and a community-based agreement on the common values for metalabels.

To overcome the deficiency of flat representation of labels in general language dictionaries, TEI Lex-0 now 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. To apply a domain label inside a sense, use the `<usg>` element with a `@corresp` attribute pointing to the `xml:id` of the appropriate category in the taxonomy.

Moreover, to overcome the above mentioned deficiency, we would ideally aim at a kind of encoding in which we can separate canonical, possibly multilingual, labels that are defined in one place and then simply pointed to from the lemma. For this reason, we propose to employ the mechanism for the definition of taxonomies already available in the `<teiHeader>`.

With this approach, domain labels are documented in `<encodingDesc>` (encoding description). The domains established in the taxonomy are declared in `<classDecl>` (classification declarations). This element is used to group the source of the domain's taxonomy used by the header or elsewhere in the document. First, the `<taxonomy>` element identifies the structured taxonomy. The categories are documented in the `<category>` element. 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 the identified language. A single category may contain more than one `<catDesc>` child, and can be described in different languages (`xml:lang`). As a result of this thought process, we can establish a multilingual hierarchy for the MATHEMATICAL SCIENCES superdomain (Fig. 7).

Flat usage label lists are usually encoded as text values of the `<usg>` element. For the sake of human readability, one could deploy the same strategy and explicitly add the domain label as the content of the `<usg>` element even when the full label taxonomy is maintained in the `<teiHeader>`. This would be particularly useful when the labels used in a given dictionary are not consistent.

This approach facilitates browsing and querying the TEI encoding of the dictionary based on the structure of the domain classification. The latter is relevant for the linked data publication of the Morais Silva dictionary data, in which the URI (Universal Resource Identifier) of each domain in the classification can be used for identifying and querying RDF data.

```
<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>
```

Fig. 7 Multilingual hierarchy of the MATHEMATICAL SCIENCES superdomain

This hierarchical organisation constitutes the foundation of the domain ontology described in the next section.

7. Domain Ontology

An important component of the MorDigital project concerns the modelling of domain ontologies covering the subject fields referred to by the domain labels of the Morais dictionary. Following the model of the ongoing work on knowledge organisation of MEDICAL AND HEALTH SCIENCES domains (see OntoDomLab-Med⁹), we propose the following organisation for GEOMETRY and ARITHMETIC – two subdomains of the MATHEMATICS domain:

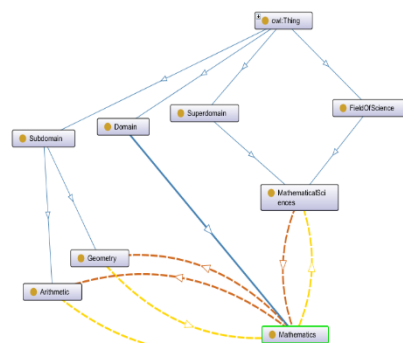


Fig. 8 Representation of the (current) knowledge organisation of MATHEMATICS, a branch of MATHEMATICAL SCIENCES, in Protégé (OntoGraf)

⁹ <https://clunl.fcsh.unl.pt/en/online-resources/ontologies/ontodmlab-med/>

Fig. 8 is a representation of the model in OWL resorting to the plugin OntoGraf¹⁰ in Protégé¹¹, a free, open-source ontology editor. This means that each class or individual in the ontology will be assigned a URI, used to reference the label present in each of the lexicographic entries in accordance – whenever possible – with the TEI schemas.

As depicted in Fig. 8, the owl:Class Mathematics relates to the owl:Class MathematicalSciences through the owl:ObjectProperty branchOf, which is represented by the yellow-dotted arc (colour publication)¹². The orange-dotted arc represents its inverse relationship, namely hasBranch. Such a decision ties in with the intent of preventing an organisation in the form of a ‘taxonomy as a subsumption-oriented hierarchy, in order to avoid misleading representations. Instead, the hasBranch relationship and its inverse branchOf relationship have been added to the Object Property hierarchy in Protégé to support this proposed knowledge organisation and facilitate subsequent linking to the lexicographic information annotated in TEI Lex-0’ (Costa et. al, 2020, p. 218).

```
<!-- http://www.semanticweb.org/OntoDomLab-Math#MathematicalSciences -->
<owl:Class rdf:about="http://www.semanticweb.org/OntoDomLab-Math#MathematicalSciences">
  <rdf:type rdfs:Class />
  <owl:intersectionOf rdfs:property="Collection">
    <rdf:Description rdf:about="http://www.semanticweb.org/OntoDomLab-Math#FieldOfScience"/>
    <rdf:Description rdf:about="http://www.semanticweb.org/OntoDomLab-Math#Superdomain"/>
    <owl:restriction>
      <owl:objectProperty rdfs:resource="http://www.semanticweb.org/OntoDomLab-Math#hasBranch"/>
      <owl:someValuesFrom rdfs:resource="http://www.semanticweb.org/OntoDomLab-Math#Mathematics"/>
    </owl:restriction>
  </owl:intersectionOf>
</owl:Class>
</rdf:subClassOf>
</owl:Class>

<!-- http://www.semanticweb.org/OntoDomLab-Math#Mathematics -->
<owl:Class rdf:about="http://www.semanticweb.org/OntoDomLab-Math#Mathematics">
  <rdf:type rdfs:Class />
  <owl:intersectionOf rdfs:property="Collection">
    <rdf:Description rdf:about="http://www.semanticweb.org/OntoDomLab-Math#Domain"/>
    <owl:restriction>
      <owl:objectProperty rdfs:resource="http://www.semanticweb.org/OntoDomLab-Math#hasBranchOf"/>
      <owl:someValuesFrom rdfs:resource="http://www.semanticweb.org/OntoDomLab-Math#MathematicalSciences"/>
    </owl:restriction>
  </owl:intersectionOf>
</owl:Class>
</rdf:subClassOf>
</owl:Class>

<owl:restriction>
  <owl:objectProperty rdfs:resource="http://www.semanticweb.org/OntoDomLab-Math#hasBranchOf"/>
  <owl:someValuesFrom rdfs:resource="http://www.semanticweb.org/OntoDomLab-Math#Mathematics"/>
</owl:restriction>
</rdf:subClassOf>
</owl:Class>

<owl:restriction>
  <owl:objectProperty rdfs:resource="http://www.semanticweb.org/OntoDomLab-Math#hasBranchOf"/>
  <owl:someValuesFrom rdfs:resource="http://www.semanticweb.org/OntoDomLab-Math#Mathematics"/>
</owl:restriction>
</rdf:subClassOf>
</owl:Class>

<owl:label xml:lang="en">Mathematics</rdf:label>
</owl:Class>
```

Fig. 9 The OWL file illustrating the formal definitions of <MathematicalSciences> and <Mathematics>

As shown in the OWL file (Fig. 9), MathematicalSciences is expressed in OWL 2 (Web

Ontology Language)¹³ and is a subclass of FieldOfScience and of Superdomain, and has (at least) Mathematics as a branch, whereas Mathematics is a subclass of Domain and has two subdomains: Geometry and Arithmetic. This assertion can be informally represented as follows:

```
MathematicalSciences is_a FieldOfScience
MathematicalSciences is_a Superdomain
MathematicalSciences hasBranch Mathematics
Mathematics is_a Domain
Mathematics hasBranch Geometry
Mathematics hasBranch Arithmetic
Geometry is_a Subdomain
Arithmetic is_a Subdomain
```

The owl:Classes Superdomain, Domain and Subdomain were added as a flat hierarchy to improve the expressivity of the ontology in order to answer the need for a classification when it comes to hierarchically organising domain labels. To express the conditions for an individual to classify as pertaining to one of the above-mentioned domain classes, we decided to define the concepts by means of intersection instead of creating an additional owl:ObjectProperty. This means that, in addition to the properties of being, on the one hand, a branch of MathematicalSciences and, on the other hand, of having 2 branches (i.e., Arithmetic and Geometry), Mathematics is also a domain.

Both owl:Class Geometry and Arithmetic are subdomains of Mathematics, expressed by the owl:ObjectProperty branchOf (represented by yellow-dotted arcs in Fig. 8 – for non-colour publications see footnote 12), and declared as *disjoint* in their formal definition through the restriction owl:disjointWith. This restriction means that an individual cannot classify as a member of those two classes simultaneously – a feature of OWL 2 that allows us to avoid ambiguity and/or inconsistency of the ontology. The consistency of the ontology, OntoDomLab-Math, is validated by one of the plugin reasoners of Protégé, namely Hermit¹⁴, which can also identify subsumption¹⁵ relationships between classes, therefore validating the correctness of our logical constructs. An example of inferred subsumption is illustrated in Fig. 9:

¹⁰ <https://protegewiki.stanford.edu/wiki/OntoGraf>

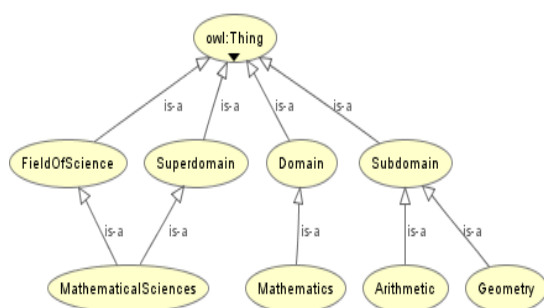
¹¹ <https://protege.stanford.edu/>

¹² For non-coloured publications: the yellow-dotted arcs are horizontally depicted with a left-right direction arrow, and with a bottom-up arrow when vertically depicted. Orange-dotted arcs are their inverse (right-left and up-bottom).

¹³ <https://www.w3.org/TR/owl-ref/>

¹⁴ <http://www.hermit-reasoner.com/>

¹⁵ Classification of concepts determines subconcept/superconcept relationships (called subsumption relationships in DL) between the concepts of a given terminology, and thus allows one to structure the terminology in the form of a subsumption hierarchy. (Baader and Nutt, 2003, p. 47).



inferred by the reasoner

The OntoDomLab-Math ontology is currently under development. Further individuals, namely domains related to MATHEMATICAL SCIENCES pointed out by domain labels, will be added once the identification of all related labels found in the Morais dictionary is concluded. We have therefore been using only one relationship beyond subsumption (SubClassOf) to define mathematics-related domain labels at the moment.

As mentioned, each class of the ontology is assigned a URI. In the case of Geometry, the URI is: <http://www.semanticweb.org/OntoDomLab-Math#Geometry> (see Fig. 9) – a unique identifier that will be declared in the TEI schema of its corresponding lexicographic article.

Concerning the encoding, the domain labels will be linked through the @corresp attribute both to the corresponding domain in the TEI-encoded classification, in the ontology of MATHEMATICAL SCIENCES and also in the SKOS version of the MorDigital domain classification.

Conclusion

This project will contribute towards a more significant presence of lexicographic digital content in Portuguese through open tools and standards.

A rigorous linguistic treatment will make it possible to organise and structure the lexicographic components, and to elicit lexical relationships between various elements.

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.

Combining semasiological and onomasiological approaches applied to the three editions of Morais will be possible via the inclusion of ontologies (e.g. diasystematic marking, namely domain labels, registers and part of speech categories).

For lexicographic content organisation, we believe it will be helpful to establish a hierarchical structure in general language dictionaries for two main reasons: 1) to organise an increasing amount of terminological information included in lexicographic resources and 2) to provide the lexicographers with greater control over specialised content in order to be able to detect inconsistencies and monitor their work more efficiently.

In a near future, we are going to organise all the domain labels – a flat list of abbreviations – that can be found in the front matter of the Morais da Silva dictionary, applying the same philosophy that underpins OntDomLab-Med and OntoDomLab-Maths, while taking into consideration the specificities of the knowledge organisation that underlies the various domains.

In conclusion, the main goal of this research is to move from flat lists (non-hierarchical lists) to hierarchical lists. Associating them to domain ontologies is a crucial step to allow interoperability between resources regardless of the languages. From a methodological point of view, applying terminological reasoning to lexicographic work has been proven to be beneficial for the sake of coherence and systematicity. In theory, the methodology we advocate can be replicated in other retrodigitised dictionaries.

Funding

This work is supported by (1) the MORDigital – Digitalização do *Dicionário da Língua Portuguesa* de António de Morais Silva [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.

References

- Almeida, B. Costa, R., Salgado, A., Ramos, M., Romary, L., Khan, F., Carvalho, S., Khemakhem, M., Silva, R., and Tasovac, T. (2022). Modelling usage information in a legacy dictionary: from TEI Lex-0 to Ontolex-Lemon. *COMHUM 2022 – 2nd Workshop on Computational Methods in the Humanities*, 9-10 June 2022, Lausanne, Switzerland.
- Baader, F., and Nutt, W. (2003). Basic Description Logics. In F. Baader, D. L. McGuinness, D. Nardi, &

- P. F. Patel-Schneider (Eds.), *The Description Logic Handbook: Theory, implementation, and applications* (2nd ed., pp. 47–100). Cambridge University Press.
- Costa, R. (2006). Texte, terme et contexte. In Blampain, D., Thoiron, P., & Van Campenhoudt, M. (Eds.), *Mots, termes et contextes. Actes des VII Journées Scientifiques du Réseau Lexicologie, Terminologie et Traduction* (pp. 79–88). Paris: Éditions des Archives Contemporaines.
- Costa, R., Carvalho, S., Salgado, A., Simões, A., and Tasovac, T. (2020). Ontologie des marques de domaines appliquée aux dictionnaires de langue générale. In Xavier Blanco (Ed.), *La lexicographie en tant que méthodologie de recherche en linguistique Revue de Philologie Française et Romane – Langue(s) & Parole*, n. 5. Mons: Edition du CIPA. pp. 201–230. ISSN 2466-7757, ISSN 2684-6691.
- Costa, R., Salgado, A., Almeida, B. (2021). SKOS as a key element for linking lexicography to digital humanities. In Koraljka Golub and Ying-Hsang Liu (Eds.), *Information Organization in Digital Humanities: A Global Perspective*. Coll. Digital Research in the Arts and Humanities, pp. 178–204. Routledge. ISBN 97803675516.
- Costa, R., Salgado, A., Kahn, F., Carvalho, A., Romary, L., Almeida, B., Khemakhem, M., Ramos, M., Silva, R., and Tasovac, T. (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.
- Costa, R., Roche, C., and Salgado, A. (2022). Standards for Representing Lexicographic Data: An Overview. <https://elexis.humanistika.org/resource/posts/standards-for-representing-lexicographic-data-an-overview>.
- Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers, etc.* Denis Diderot and Jean le Rond d'Alembert (Eds.). University of Chicago: ARTFL Encyclopédie Project (Autumn 2022 Edition), <http://encyclopedia.uchicago.edu/>.
- ISO 1087 (2019). Terminology Work – Vocabulary – Part 1: Theory and Application. Geneva: International Organization for Standardization.
- ISO 704 (2022). Terminology work – Principles and methods. Geneva: International Organization for Standardization.
- Khan, F., Romary, L., Salgado, A., Bowers, J., Khemakhem, M., and Tasovac, T. (2020). Modelling Etymology in LMF/TEI: The 'Grande Dicionário Houaiss da Língua Portuguesa' Dictionary as a Use Case. In N. Calzolari et al. (Eds.), *LREC 2020 Conference Proceedings*. Paris: ELRA, pp. 3172–3180. ISBN 979-10-95546-34-4.
- Khemakhem, M., Galleron, I., Williams, G. Romary, L., and Suárez, P. J. O. (2019). How OCR Performance Can Impact on the Automatic Extraction of Dictionary Content Structures. In *19th Annual Conference and Members' Meeting of the Text Encoding Initiative Consortium*. Austria: Graz. <https://hal.archives-ouvertes.fr/hal-02263276>.
- Khemakhem, M., Foppiano, L., and Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources Using Conditional Random Fields. In *Proceedings of eLex 2017 Conference: Electronic lexicography in the 21st century: Lexicography from Scratch*. Netherlands: Leiden, pp. 598–613.
- Morais Silva, A. M. (1789). *Diccionario da lingua portugueza composto pelo padre D. Rafael Bluteau, reformado, e accrescentado por Antonio de Moraes Silva, natural do Rio de Janeiro* (Vol. 1–2). Officina 730 de Simão Thaddeo Ferreira. <https://purl.pt/29264>.
- Romary, L., Khemakhem, M., Khan, F., Bowers, J., Calzolari, N., et. al. (2019). LMF Reloaded. *AsiaLex 2019: Past, Present and Future*, June 2019, Istanbul, Turkey. Ffhal-02118319f.
- Sager, J. C. (1990). *A practical course in terminology processing*. Amsterdam: John Benjamins Publishing Company.
- Salgado, A. (2021). *Terminological Methods in Lexicography: Conceptualising, Organising and Encoding Terms in General Language Dictionaries*. Doctoral dissertation. Universidade NOVA de Lisboa. <https://run.unl.pt/handle/10362/137023>.
- Salgado, A., and Costa, R. (2019). Marcas temáticas en los diccionarios académicos ibéricos: estudio comparativo. *RILEX: Revista sobre investigación léxicos*, 2(2), pp. 37–63. e-ISSN 2605-3136.
- Salgado, A., Costa, R., and Tasovac, T. (2019). Improving the consistency of usage labelling in dictionaries with TEI Lex-0. *Lexicography: Journal of ASIALEX*. e-ISSN 2197-4306.
- Salgado, A., Costa, R., and Tasovac, T. (2022). Applying terminological methods to lexicographic work: Terms and their domains. In A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs, and P. Storjohann (Eds.), *Dictionaries and Society. Proceedings of the XX EURALEX International Congress* (pp. 181–195). IDS-Verlag.
- Silva, R. (2014). *Gestão de terminologia pela qualidade. Faculdade de Ciências Sociais e Humanas*. Doctoral dissertation. Universidade NOVA de Lisboa. <http://hdl.handle.net/10362/13664>.
- Svensén, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge University Press.
- Tasovac, T., Romary, L., Banski, P., Bowers, J., Does, J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Lehečka, B., Petrović, S., Salgado, A., and Witt, A. (2018). *TEI Lex-0: A baseline encoding for lexicographic data*. Version 0.9.0. DARIAH Working Group on Lexical Resources. <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>.