Modelling usage information in a legacy dictionary: from TEI Lex-0 to Ontolex-Lemon

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

Modelling Usage Information in a Legacy Dictionary: From TEI Lex-0 to Ontolex-Lemon

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

[1]*NOVA CLUNL - Centro de Linguística da Universidade Nova de Lisboa, Portugal*
[2]*ACL - Academia das Ciências de Lisboa, Portugal*
[3]*Inria - Institut national de recherche en sciences et technologies du numérique, France*
[4]*ILC-CNR - Istituto di Linguistica Computazionale "Antonio Zampolli", Italy*
[5]*CLLC - Centro de Línguas, Literaturas e Culturas da Universidade de Aveiro, Portugal*
[6]*ArcaScience, France*
[7]*BCDH - Belgrade Center for Digital Humanities, Serbia*

Abstract

This paper describes ongoing work in the modelling of usage information in the context of the MORDigital project. The latter is based on the encoding and publication as linked data of *Diccionario da Lingua Portugueza*, a Portuguese legacy dictionary authored by António de Morais Silva, whose first edition was published in 1789. In this paper, we focus on the TEI Lex-0 encoding and Ontolex-Lemon modelling of lexicographic articles from the Morais Silva dictionary that feature usage information. The approach described in this paper should be reusable for other projects involving the encoding and linked data publication of legacy dictionaries.

Keywords

legacy dictionaries, usage information, lexicography, digital humanities

COMHUM 2022: Workshop on Computational Methods in the Humanities, June 09–10, 2022, Lausanne, Switzerland

*Corresponding author.

†These authors contributed equally.

brunoalmeida@fcsh.unl.pt (B. Almeida); rute.costa@fcsh.unl.pt (R. Costa); anasalgado@fcsh.unl.pt (A. Salgado); mvramos@fcsh.unl.pt (M. Ramos); laurent.romary@inria.fr (L. Romary); fahad.khan@ilc.cnr.it (F. Khan); sara.carvalho@ua.pt (S. Carvalho); mohamed.khemakhem@inria.fr (M. Khemakhem); raq.silva@fcsh.unl.pt (R. Silva); ttasovac@humanistika.org (T. Tasovac)

0000-0002-5777-5574 (B. Almeida); 0000-0002-3452-7228 (R. Costa); 0000-0002-6670-3564 (A. Salgado); 0000-0001-7209-3806 (M. Ramos); 0000-0002-0756-0508 (L. Romary); 0000-0002-1551-7438 (F. Khan); 0000-0002-7501-5405 (S. Carvalho); 0000-0003-3529-2990 (M. Khemakhem); 0000-0002-0505-4863 (R. Silva); 0000-0002-3919-993X (T. Tasovac)

1. Introduction

The publication of *Diccionario da Lingua Portugueza* in 1789, authored by António de Morais Silva, marks the beginning of contemporary Portuguese lexicography, following the model set by several modern language dictionaries published in Europe in the 17th and 18th centuries [1, 2]. As the first Portuguese monolingual dictionary, it had a fundamental role in the standardisation of this language, and constitutes a reference for studying the evolution of the Portuguese lexicon [3]. The first edition of the dictionary had two volumes (Vol. 1, 752 p. and Vol. 2, 541 p.). Morais directly oversaw the 2nd and 3rd editions (published, respectively, in 1813 and 1823). This work was greatly revised and updated over the years, culminating in the 10th edition, which was published in 12 volumes from 1949 to 1959.

The MorDigital project[1] aims at digitising and publishing in open access the first three editions of the dictionary by Morais [4]. Our methodology involves the reuse of digitised versions of the dictionary, available in the public domain as PDF files. The latter are currently undergoing a re-OCRisation process to ensure the quality of the final output of the project. The digitised versions of the dictionary will be structured by means of several open standards for encoding and modelling lexical and dictionary data, which will facilitate interoperability with existing systems and datasets.

The encoding of the dictionary's editions will be carried out in TEI Lex-0 [5], a baseline XML encoding for machine-readable dictionaries based on the guidelines of the Text Encoding Initiative (TEI). The TEI Lex-0 encoding of the Morais Silva dictionary will be the basis for a LMF (Lexical Markup Framework) version, which should be facilitated by the ongoing convergence between TEI and the LMF standard [6]. The TEI Lex-0 encoding of the Morais Silva dictionary will be further transformed to RDF based on Ontolex-Lemon [7], a model originally developed for enriching ontologies with lexical information, which has become a *de facto* standard for publishing lexical resources as linked data [8]. The recently developed lexicography module, *lexicog* [9], facilitates the application of Ontolex to dictionary data. An XSLT-based tool, `tei2ontolex`[2], will be used for the conversion process. Examples such as those presented in this paper will be the basis for a wider coverage of features of this tool.

While the work shown in this paper is focussed on digital lexicography, specifically on the retrodigitisation of legacy dictionaries, technologies such as TEI and Ontolex are relevant in many other domains of the digital humanities involving text encoding, analysis, and publishing, including discourse analysis and other fields of linguistics, digital literary studies, cultural heritage, and digital archives. TEI, for instance, includes several communities of practice whose activity is centred on encoding text in a standardised, flexible, and interoperable framework. This paper could, therefore, be useful for scholars in those communities, especially for those whose work also involves applying linked data models for publishing digital humanities resources.

[1] https://mordigital.fcsh.unl.pt/

[2] https://github.com/elexis-eu/tei2ontolex

1.1. Related work

Relevant work for the approach described in this paper has been carried out recently. Much of this work is focussed on relating TEI Lex-0 with Ontolex-Lemon, augmented with other ontologies, for the linked data modelling of lexicographic resources, emphasising applications in retrodigitisation projects[3]. The importance of the above-mentioned formats and models is made clear in the context of the ELEXIS project[4], a European infrastructure for interoperable lexicographic resources, in which TEI Lex-0 and Ontolex-Lemon are two of the main formats for publishing and interlinking dictionary data [12].

Khan and Salgado [13] describe a novel approach to the modelling and publication of lexicographic resources as linked data. This approach consists of using Ontolex-Lemon and lexicog in conjunction with the CIDOC-CRM aligned FRBRoo ontology for representing different levels of description of lexicographic resources (i.e., work, expression, manifestation) corresponding to the different views of dictionaries explained in the TEI Guidelines [14], namely typographical (the layout of the pages), editorial (the text of the dictionary) and lexical (the conceptual and linguistic content of the dictionary). The work carried out in this paper pertains to the lexical view of the Morais dictionary, in which elements from Ontolex and lexicog are used to model usage information following an interoperable approach to that of Khan and Salgado [13] (see Section 6 of this paper).

In addition to the work described above, current research has focussed on lexicography and digital humanities, including the application of ontologies and knowledge organisation. Costa et al. [15] show how domain labels can be modelled through an OWL ontology in the medical and health sciences (OntoDom-Lab-Med[5]), whose classes can be applied to the semantic annotation of usage information in TEI Lex-0 encoded dictionaries. This can be done solely with TEI elements and attributes, including the ontology class URI within the usage information element, or with the XML Linking Language (XLink), which also allows to describe the role played by ontology class URI, providing more complex domain information.

In turn, Costa et al. [16] show how SKOS (Simple Knowledge Organization System), a W3C recommendation for modelling knowledge organisation schemes, can be employed in digital humanities projects for modelling linguistic/lexicographic categories represented in dictionaries' lists of abbreviations (e.g., part of speech, grammatical gender, register). SKOS modelling, for knowledge organisation purposes, is shown to be complementary to the TEI Lex-0 encoding of dictionary articles.

Salgado et al. [17] further highlight the importance of domain label modelling through terminological methods, namely by structuring domain labels. The resulting taxonomies or classifications of domains can be included directly within the `<teiHeader>` element of TEI-encoded dictionaries, whose categories can be linked to individual dictionary articles through the TEI usage information element, while still retaining the text values that occur in the dictionary articles for human readability purposes.

[3] See for example [10, 11]

[4] https://elex.is/

[5] https://doi.org/10.34619/emw4-ax6o

Section 5 of this paper illustrates the simpler option, as described in Costa et al. [15], for linking the TEI encoding of usage information to OntoDomLab-Med and to the MorDigital Domain Classification. The latter, described in Section 4, follows notions laid out in Costa et al. [16] with regard to the complementarity between TEI encoding and SKOS modelling, and Salgado et al. [17] regarding the application of terminological methods in lexicography for structuring domain labels.

2. Background

2.1. Usage information in lexicography

In lexicographic theory, usage or diasystematic[6] information is understood as a set of constraints or restrictions on the use of words, or their senses, to certain contexts or to a subset of language users (e.g., [19, 20]). Dictionaries traditionally include usage information in the lexicographical articles as labels (often abbreviated), or in more verbose forms, such as notes or as part of the lexicographic definitions themselves. As Svensén notes, diasystematic marking in lexicographic articles implies that "a certain lexical item deviates in a certain respect from the main bulk of items described in a dictionary" [20, p. 315]. This notion of deviation from the lexicon of standard varieties of languages is, therefore, the cornerstone of usage marking in dictionaries.

As noted by Salgado et al., usage labels are devices whose simplicity is only apparent, since they "often conceal the complexity of dynamic sociolinguistic, cultural, and ideological processes that they are meant to illustrate" [21, p. 134]. In a digital humanities context, such as the retrodigitisation of dictionaries, labels are also a challenge for the interoperability of lexicographic datasets from different sources, ranging over different cultures, languages, and periods. Indeed, since at least the 17th century, lexicographers have included labelling/marking in dictionaries, describing a wide variety of usage information, whose treatment in theoretical and practical lexicography has not always been consistent [22].

There have been several surveys of usage information in theoretical and practical lexicography, such as Ptaszyński [22] and Vrbinc and Vrbinc [23]. A more recent and comprehensive survey has been carried out by Salgado et al. [21]. These studies note the difficulties caused by the variety, and partial overlapping, of usage information types put forward by lexicographers, and the different terminology employed by them. Hausmann [24] put forward the most comprehensive classification, with 11 types of usage information. This classification was later adopted by Bergenholtz and Tarp [25] and Svensén [20]. Landau [19], whose manual was first published in the same year as Hausmann's proposal, distinguished 9 types of usage information. In a later study, Milroy and Milroy [26] included 5 types of usage information, distinguishing 'group labels', which pertain to a subset of language users, from 'register labels', pertaining to specific social and communicative contexts. Jackson [27] put forward a classification including 7 types of usage information. Atkins and Rundell [28] considered 9 types of usage information, which they call 'linguistic labels'.

[6]The term 'diasystem', originating from dialectology [18], designates a general language system encompassing several dialects. In the context of lexicography, the term 'diasystematic' applies to the marking of lexical items whose usage *deviates* from that of the lexicon of the standard variety of a language.

Table 1: Usage information typologies in lexicographic literature

Hausmann (1989) adopted by others	Milroy, J. and Milroy, L. (1990)	Landau (2001)	Jackson (2002)	Atkins and Rundell (2008)	TEI Lex-0 usage type	Criterion	Examples
diachronic	temporal	currency/temporality	history	time	temporal	Time	archaic, old
diatopic	geographical	regional/geographical variation	dialect	region/dialect	geographic	Place	AmE., dial.
diaintegrative	—	—	—	—	hint	Nationality	Latin, English
diamedial	—	style, functional variety/register	—	—	hint	Medium	spoken
diastratic	—	restricted or taboo sexual scatological usage and slang	status	slang and jargon/offensive terms	socioCultural	Sociocultural	slang, vulgar, formal
diaphasic	register	style, functional variety/register	formality	register	socioCultural	Formality	slang, vulgar, formal
diatextual	—	style, functional variety/register	—	style	textType	Text type	bibl., poet., admin., journalese
diatechnical	field	technical or specialised terminology	topic or field	domains	domain	Speciality	Med., Biol., Phys.
diafrequential	frequency	—	—	—	frequency	Frequency	rare, occas.
diaevaluative	—	insult/style, functional variety/register	effect	attitude	attitude	Attitude	derog., euph.
dianormative	—	status or cultural level	disputed usage	—	normativity	Normativity	non-standard, incorrect
—	—	—	—	—	meaningType	Meaning	fig., lit.

Table 1 shows a systematisation of the usage types considered in the above-mentioned classifications, based on the survey carried out by Salgado et al. [21].

2.2. TEI P5 and TEI Lex-0

The Text Encoding Initiative, or TEI, is an international organisation with a long history in the development of guidelines, and associated schemas, for encoding machine-readable text in social sciences and humanities. The current release of the guidelines, TEI P5, include a module (i.e., a set of XML elements and attributes) for encoding dictionaries and other lexical resources, e.g., glossaries and word lists included in other documents [14, 9: Dictionaries]. The common characteristic of these resources is that they consist of entries/articles describing lexical items in a language (or languages). Since the characteristics of these resources may vary widely, TEI P5 includes the `<entry>` element for encoding conventional dictionary articles, the `<entryFree>` element for encoding unstructured entries in generic lexical resources, and the `<superEntry>` element for grouping several lexical entries.

In a straightforward TEI encoding of print dictionaries, the main body of text (encoded through the `<body>` element) should include a set of `<entry>` elements, corresponding to the dictionary's articles. Among other possibilities, each `<entry>` element may contain:

- Information about written and/or spoken forms of the headword (through the `<form>` element).
- Grammatical information (within the `<gramGrp>` element).
- Information about the headword's senses (through the `<sense>` element).
- Cited quotations (through the `<cit>` element, part of the core TEI elements).
- Usage information (through the `<usg>` element).

As defined in TEI P5, the <usg> element may appear in several components and positions of the lexicographic article structure (`<entry>` element). The `<usg>` element has an optional `@type` attribute for indicating types of usage information, and the guidelines include 16 sample values, which have been adopted in many projects.

The flexibility of TEI P5, allowing numerous possibilities and combinations of elements for encoding dictionaries, is a hurdle for interoperability of dictionary data emanating from different projects and for its use by NLP applications. TEI Lex-0 is a more recent initiative at establishing a baseline encoding and target format for TEI dictionaries [5]. It introduces a number of constraints on the encoding of lexicographical articles, such as limiting the possible occurrences of the `<def>` element (used for definition texts) to `<sense>`, while in TEI P5 `<def>` could appear directly within `<entry>` and other elements.

With regard to usage information, in TEI Lex-0 the <usg> element may still occur in several points of the entry hierarchy, but the `@type` attribute is made mandatory. A more concise list of 10 values for possible usage types was also introduced, following Salgado et al. [21][7]:

[7]The TEI Lex-0 reference document includes a table showing the correspondence between the suggested values of `usg/@type` in TEI P5, their required values in TEI Lex-0 and some examples of real dictionary data [5, sec. 8.2. Types of usage].

- `"temporal"`. Marks the usage of a lexical item in a scale from old to new (this is known as **diachronic information** in lexicographic literature).
- `"geographic"`. Marks the place or region where a lexical item is mostly used (**diatopic information**).
- `"domain"`. Marks the subject field in which the lexical item is mostly used (**diatechnical information**).
- `"frequency"`. Marks the relative occurrence of a lexical item (**diafrequential information**).
- `"textType"`. Marks the typical discource type or genre where a lexical item is mostly used (**diatextual information**).
- `"attitude"`. Marks the speaker's subjective point of view regarding the referent of a lexical item (**diaevaluative information**).
- `"socioCultural"`. Marks the social groups (**diastratic information**) and/or communicative situations (**diaphasic information**) where a lexical item mostly occurs.
- `"meaningType"`. Marks a semantic extension of the sense of a lexical item[8].
- `"normativity"`. Marks the usage of a lexical item as non-standard or incorrect (dianormative information).
- `"hint"`. Marks a non-specified usage of a lexical item (default value of `<usg>`).

TEI Lex-0 effectively restricts the scope of `<usg>` to put it more in line with lexicographic theory, deprecating, for example, the encoding of lexical relations and etymological information through the `<usg>` element, which are both allowed in TEI P5.

2.3. Ontolex-Lemon, lexicog and LexInfo

The Lexicon Model for Ontologies (Ontolex-Lemon) was put forward by the W3C Ontology-Lexicon community group for the enrichment of ontologies with linguistic information in the Semantic Web [7]. This model has since become a *de facto* standard for modelling lexical resources as RDF and publishing them as linguistic linked data [8]. Ontolex-Lemon was heavily inspired by the core model of the Lexical Markup Framework (LMF), an ISO standard for machine-readable lexical resources [29], having transposed several LMF classes for modelling linguistic information. These include the following:

- `LexicalEntry`. A lexical entry is a unit of a lexicon consisting of a set of grammatically related forms associated with a collection of senses (e.g., cat in English, including both singular and plural forms, which are associated with several senses in this language).
- `Form`. A form is a grammatical realisation of a lexical entry (e.g., the singular form 'cat', which is the lemma or canonical form for representing the entry).
- `LexicalSense`. A sense associated with a lexical entry, which can be described, e.g., in a dictionary definition.

[8]Labels for marking the semantic extension of senses (e.g., 'figurative', or 'fig.') are not considered in the typologies of usage information of lexicographic literature. Nevertheless, it remains a possible usage type in TEI Lex-0, maintaining interoperability with the `style` usage type of TEI P5.

- `Lexicon`. A lexicon is a collection of lexical entries for a particular language.

While the core model has enough elements to describe information about the lexicon of individual languages, it lacks expressive power to properly describe *lexical resources*, such as dictionaries. Indeed, lexicographic articles often include information about forms shared by different parts of speech, which necessarily correspond to different lexical entries in Ontolex-Lemon. *Lexicog*, the Ontolex-Lemon Lexicography Module, aims to address these issues, based on several experiences in converting to linked data existing lexicographic resources [9]. *Lexicog* introduces classes and properties that enable the distinction of the lexicon and its lexicographic description. The following are the most relevant classes of *lexicog*:

- `Entry`. An entry is an element of a dictionary's microstructure, corresponding to a lexicographic article.
- `LexicographicComponent`. A lexicographic component is an element for describing substructures of dictionary entries (e.g., senses, sense groups or subentries in a lexicographic article).
- `LexicographicResource`. A lexicographic resource is a collection of lexicographic articles.

With these elements, it becomes possible to distinguish between *lexical entries* (which must belong to the same part of speech, such as noun or adjective) and *dictionary entries* (in which different parts of speech may be conflated). Sense groupings and subentries can be modelled as lexicographic components, which in turn describe either lexical entries or lexical senses in the language's lexicon.

The means for modelling usage information fall within the core Ontolex-Lemon model, which includes the `usage` property. The latter allows to represent modulations in the meaning of lexical entries determined by "usage conditions or pragmatic implications" [7], such as due to register, connotations, etc. The domain of `usage` is restricted to instances of `LexicalSense`. This is an important distinguishing feature of Ontolex-Lemon when compared to the TEI abstract model: while in the latter usage information may appear directly in several parts of an entry, in Ontolex-Lemon usage is necessarily associated with lexical senses.

While Ontolex-Lemon includes the `usage` property, it does not specify how to represent usage information. Furthermore, Ontolex-Lemon does not include elements for representing *types* of usage information, or other linguistic categories (e.g., part of speech). Ontolex-Lemon relies on external vocabularies for describing the properties of linguistic objects. The LexInfo ontology[9], in particular, was created to provide linguistic categories for modelling in Ontolex-Lemon. The former declares 13 sub-properties of the Ontolex-Lemon `usage` property, which adopt the above-mentioned usage types of TEI Lex-0[10].

[9] http://lexinfo.net/

[10] LexInfo and Ontolex-Lemon include three additional sub-properties of `usage`, (`condition`, `normativeAuthorization` and `register`) but they pertain to domains of application other than lexicography, namely argument structure and terminological databases.

3. Usage information in the Morais Silva dictionary

3.1. Typology of usage information

In the Morais Silva dictionary, usage information is mostly marked by abbreviated labels, the full form of which is given in the dictionary's explanation of abbreviations[11]. The analysis of the latter provides valuable insights into the usage information that was relevant to the late 18th century lexicographer. Our analysis resulted in the following typology of labels (the corresponding types of usage in TEI Lex-0 and LexInfo, are shown in parentheses):

- **Diatechnical information** (`"domain"`). E.g., *Med.*, for medical terms).
- **Diatextual information** (`"textType"`). E.g., *Poet.*, for poetic words.
- **Diastratic information** (`"socioCultural"`). E.g., *Vulg.*, for words associated with the common people.
- **Diaphasic information** (`"socioCultural"`). E.g., *Fam.*, for words used in a familiar context.
- **Diatopic information** (`"geographic"`). E.g., *Asiat.*, for words used in the former Portuguese colonies in India.
- **Diachronic information** (`"temporal"`). E.g., *Ant.*, for antiquated words.
- **Diaintegrative information** (`"hint"`). E.g., *Lat.*, for Latin words integrated in Portuguese.
- **Diafrequential information** (`"frequency"`). E.g., *P. us.*, for rarely used words.
- **Semantic extension information** (`"meaningType"`). E.g., *f.* and *fig.*, marking figurative usages of lexical items).

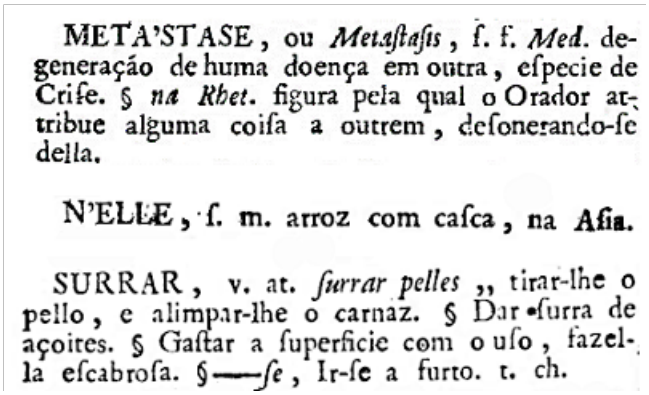
3.2. Examples of usage information in the Morais Silva dictionary

Figure 1 shows three lexicographic articles in which usage information occurs. The article *metástase* ('metastasis') [30, vol. 2, p. 79] is a straightforward example of usage information associated with word senses. The senses, separated by the section sign (§), are marked as belonging to different subject fields, namely medicine (*Med.*) and rhetoric (*na Rhet.*), constituting diatechnical information.

The article *nélle* (Morais Silva, 1789, vol. 2, p. 113) is an example of a dictionary article in which usage information is not in abbreviated form. This article includes a single sense including the parenthetical phrase *na Asia* ('in Asia'), which is diatopic information restricting the usage of the word to the speakers from the Portuguese colonies of India at the time.

The article *surrar* [30, vol. 2, p. 445] describes a verb whose base meaning is roughly equivalent to that of the English verb '(to) flesh', i.e. to remove the flesh adhering to a skin or hide. The article includes four senses, also separated by the section sign. The first sense is

[11] Although the listed abbreviations did not change significantly in the first three editions of the dictionary, it should be noted that these abbreviations are not always used in the dictionary's articles, in which the relevant information is often marked using full forms or non-listed abbreviations. For example, 'f.' is explained as an abbreviation of 'femenino' (the feminine grammatical gender), although in some articles 'f.' is also used to mark figurative senses. 'Fig.' is also used, although it is not listed in the explanation of abbreviations.



META'STASE, ou *Metaſtaſis*, ſ. f. *Med.* degeneração de huma doença em outra, eſpecie de Criſe. § *na Rhet.* figura pela qual o Orador attribue alguma coiſa a outrem, deſonerando-ſe della.

N'ELLE, ſ. m. arroz com caſca, na Aſia.

SURRAR, v. at. *ſurrar pelles* ,, tirar-lhe o pello, e alimpar-lhe o carnaz. § Dar ſurra de açoites. § Gaſtar a ſuperficie com o uſo, fazella eſcabroſa. §——*ſe*, Ir-ſe a furto. t. ch.

Figure 1: Lexicographic articles from the Morais Silva dictionary

associated to what can be construed as a citation example, *surrar peles* ('to flesh hides'). The last sense corresponds to the pronominal form of the verb, *surrar-se*, which means 'to run away', 'to remove oneself'. This sense is marked with usage information (*t. ch.* or *termo chulo* in its expanded form) indicating an informal communicative situation in which the participants joke or mock one another or a third party, which can be thought of as diaphasic information[12]. As we can see from these examples, the labels may appear at different points of the lexicographic article.

4. The MorDigital classification of domain labels

Our approach pays special attention to the encoding and modelling of domain labels in lexicographic resources, following previous work (cf. Section 1.1). This is justified in our approach by the fact that domain labels constitute an interface between lexicography and terminology, which is an important part of the interdisciplinary framework of the MorDigital project. In this context, the following complementary approaches are possible:

- Declare a taxonomy/classification of the domains referred to by the dictionary labels in the `<teiHeader>` element of the TEI Lex-0 encoding of the Morais Silva dictionary.
- Model the domain classification independently by means of W3C-maintained technologies, such as RDF, OWL and SKOS.

While the former 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 of each domain in the classification can be used for identifying and querying RDF data. Furthermore, the standalone modelling and linked data publication of the domain classification facilitates its reuse, enabling, e.g., the alignment with other KOS and domain classifications from other e-lexicography projects.

[12]The adjective 'chulo', to which the *ch.* abbreviation corresponds, is defined in the Morais Silva dictionary as "being used in familiar conversation, joking, mocking or talking fresh, as they say" [30, vol. 1, p. 170].

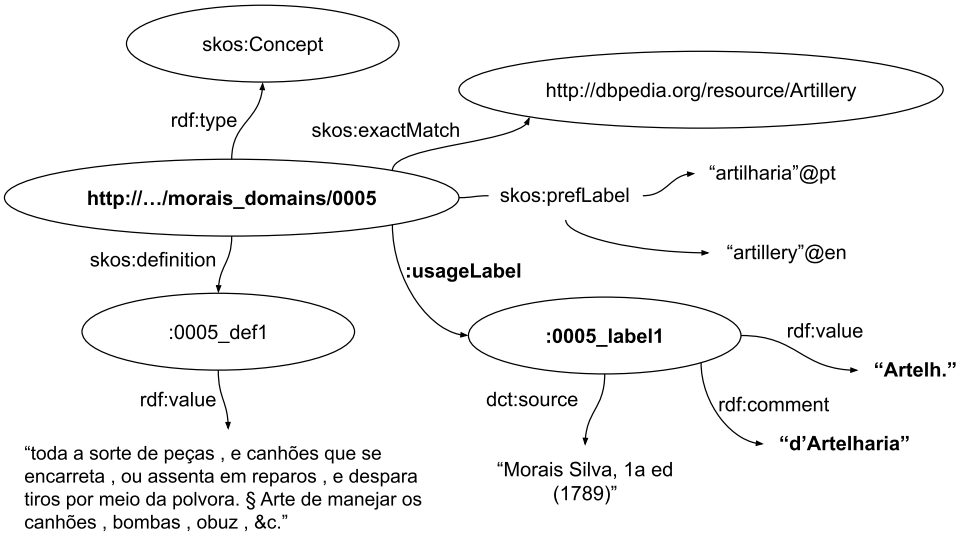


Figure 2: Example of an individual domain in the MorDigital Domain Classification

The list of domain labels appears in the front matter of the dictionary. To overcome the deficiency of flat representation of labels in general-language dictionaries, TEI-Lex 0 now recommends a hierarchical representation, an approach described in Salgado et al. [17], in which the classification is included in the `<teiHeader>` element, as previously mentioned.

The standalone modelling of the domain classification should be carried out in SKOS, the W3C model for classifications and other KOS. This model has the advantage of allowing for the straightforward modelling of hierarchies and associated networks of concepts, which can be documented with notes, designated by lexical labels (either abbreviated or in full form) and aligned with external KOS. The SKOS model can also easily be expanded with further classes and properties for expressing more information[13]. The development of this SKOS classification will be carried out in conjunction with the modelling of domain ontologies for knowledge representation in the domains included in the dictionary's list of abbreviations, of which OntoDomLab-Med is the reference for medical and health sciences.

A model is being developed for the consistent representation of information about each domain in the SKOS classification. Figure 2 shows how the model is applied to the domain of artillery (*artilharia*). In this example, we can see information about the designations of the domain in the dictionary, most importantly the label(s) that occur in the lists of abbreviations through the `:usageLabel` property. Both the abbreviated form and the full form, as explained in the dictionary's list of abbreviations, are included through standard RDF properties. This example also shows the definition of *artilharia* taken from the dictionary's first edition (definitions from the other editions would also be included).

[13]The use of SKOS as an underlying model for the classification has the advantage of leveraging the relationship with NOVA FCSH (School of Social Sciences and Humanities of NOVA University of Lisbon) through a controlled vocabulary repository it manages in the context of the ROSSIO Infrastructure for social sciences, arts, and humanities (http://vocabs.rossio.fcsh.unl.pt/).

```
<entry xmlns="http://www.tei-c.org/ns/1.0"
xml:id="MORAIS.1.DLP.METASTASE.s"
type="mainEntry" xml:lang="pt">
    <form type="lemma">
        <orth>METÁSTASE</orth>
        <metamark function="lemmaDelimiter">,</metamark>
        <metamark function="variantDelimiter">ou</metamark>
    </form>
    <form type="variant">
        <orth rend="italic">Metaſtaſis</orth>
    </form>
    <metamark function="lemmaDelimiter">,</metamark>
    <gramGrp>
        <gram type="pos" norm="NOUN">ſ.</gram>
        <gram type="gen">f.</gram>
    </gramGrp>
    <sense xml:id="MORAIS.1.DLP.METASTASE.s.1">
        <usg type="domain" corresp="#domain.medicine">Med.</usg>
        <def>degeneração de huma doença em outra,
            eſpécie de Criſe</def>
        <pc>.</pc>
    </sense>
    <metamark function="senseDelimiter">§</metamark>
    <sense xml:id="MORAIS.1.DLP.METASTASE.s.2">
        <usg type="domain" corresp="#domain.rhetoric">
            na Rhet.</usg>
        <def>figura pela qual o Orador attribue alguma coisa
            a outrem , deſonerando-se della</def>
        <pc>.</pc>
    </sense>
</entry>
```

Figure 3: *Metástase* article in TEI Lex-0

5. TEI Lex-0 encoding of usage information in lexicographic articles

Figure 3 shows the encoding in TEI Lex-0 of the *metástase* article, given as an example in Section 3.2. As explained above, usage information in TEI Lex-0 is encoded through the `<usg>` typed element, with the value `"domain"` in the case of domain labels. The actual labels for the domains, as they appear in the article, are then included as the content of the `<usg>` element, in this example *Med.* and *na Rhet.*

In this example, the labels are linked through the `@corresp` attribute to the corresponding domain in the TEI-encoded taxonomy, i.e., `#domain.medicine` and `#domain.rhetoric`. Links of these domains to external KOS and ontologies are carried out within the TEI header, where the taxonomy of domains is encoded.

Figure 4 shows a section of the taxonomy pertaining to the domains of medicine and rhetoric. Based on the results of terminological work already carried out, medicine is nested within the superdomain of medical and health sciences, while rhetoric has yet to be structured, pending the results of further terminological analysis. The `@valueDatcat` attribute is used to align the domains of the taxonomy to classes of the OntoDomLab-Med ontology of med-

```
<taxonomy xml:id="domains">
    <category xml:id="domain.medical_and_health_sciences"
        valueDatcat="http://www.semanticweb.org/OntoDomLab-Med#MedicalAndHealthSciences
http://vocabs.rossio.fcsh.unl.pt/morais_domains/0037">
        <catDesc xml:lang="en">
            <term>Medical and Health Sciences</term>
        </catDesc>
        <catDesc xml:lang="pt">
            <term>Ciências Médicas e da Saúde</term>
        </catDesc>
        <category xml:id="domain.medicine"
            valueDatcat="http://www.semanticweb.org/OntoDomLab-Med#Medicine
http://vocabs.rossio.fcsh.unl.pt/morais_domains/0025">
            <catDesc xml:lang="en">
                <term>Medicine</term>
            </catDesc>
            <catDesc xml:lang="pt">
                <term>Medicina</term>
            </catDesc>
        </category>
    </category>
    <category xml:id="domain.rhetoric"
        valueDatcat="http://vocabs.rossio.fcsh.unl.pt/morais_domains/0033">
        <catDesc xml:lang="en">
            <term>Rhetoric</term>
        </catDesc>
        <catDesc xml:lang="pt">
            <term>Retórica</term>
        </catDesc>
    </category>
</taxonomy>
```

Figure 4: Domain taxonomy within the TEI header

ical and health sciences (`http://www.semanticweb.org/OntoDomLab-Med#Medicine` and `http://www.semanticweb.org/OntoDomLab-Med#MedicalAndHealthSciences`) and concepts of the MorDigital Domain Classification in SKOS (e.g., `http://vocabs.rossio.unl.pt/morais_domains/0025` for medicine).

This approach allows for a more straightforward TEI encoding of dictionary articles, since the alignment to external KOS and ontologies is centralised within the header element, requiring less information in the articles themselves. The role of external KOS and ontologies remains essential as the end results of terminological work. This enables both the hierarchical structuring of the domain taxonomy and a richer conversion from TEI to linked data, as we will see in the following section.

6. Ontolex-Lemon modelling of usage information in lexicographic articles

The core Ontolex-Lemon model already provides most of the necessary elements for modelling information associated with lexical entries. The lexicographic module, *lexicog*, includes additional elements for information pertaining to lexicographic articles. As we saw in Section 2.3, the LexInfo ontology provides data categories for Ontolex, including several usage sub-properties aligned with TEI Lex-0 (e.g., `"domain"`, `"socioCultural"`). Finally, the domain classification in SKOS allows to organise the subject fields corresponding to the domain labels and align them

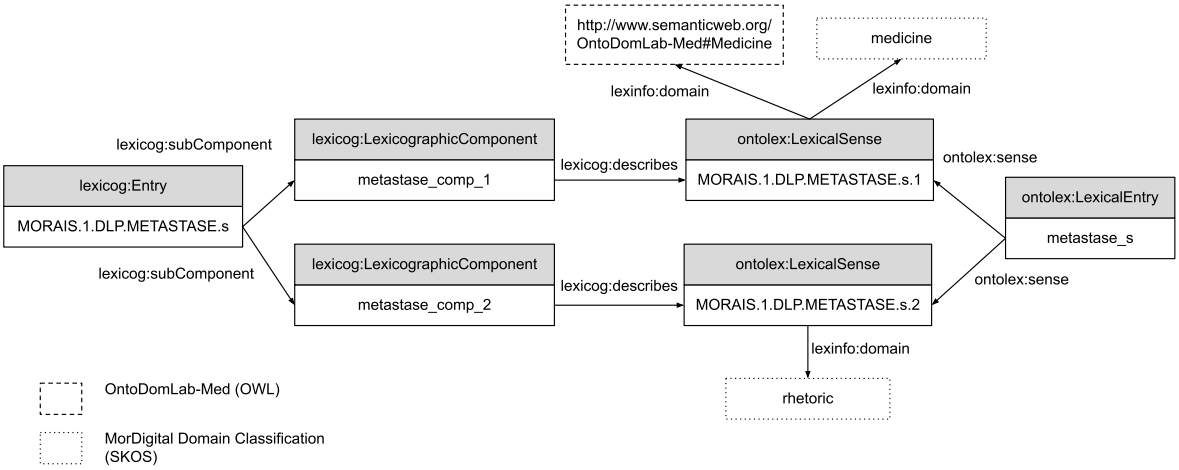


Figure 5: Usage information for the *metástase* article in Ontolex-Lemon and *lexicog*

with external knowledge bases and KOS.

Figure 5 shows the *metástase* article modelled in Ontolex, along with elements from the above-mentioned ontologies[14]. The senses are the key structural components of the article. As such, they are modelled as lexicographic components, instances of *lexicog*'s respective class. This allows to distinguish between the dictionary article and the *metástase* noun that it describes. The noun's senses are constrained to different domains (medicine and rhetoric), through the `lexinfo:domain` object property. These domains are modelled as classes/concepts in external KOS and ontologies, such as the OntoDomLab-Med ontology for medical and health sciences and the MorDigital Domain Classification, whose URI are present in the taxonomy of domains encoded in the TEI header, as seen in the previous section.

7. Future work

This paper provided an overview of usage information in lexicography and its respective encoding, through TEI Lex-0, and linked data modelling by means of Ontolex-Lemon and related ontologies. The work described in this paper is relevant in the context of the TEI-encoding and linked open data publishing of legacy dictionaries in the context of a digital humanities project.

Examples of lexicographic articles of the Morais Silva dictionary with usage information were provided, along with ongoing work in the encoding and modelling of this information. Further examples are being worked on to facilitate the conversion from the TEI Lex-0 encoding to Ontolex-Lemon by means of a XSLT-based tool, such as *tei2ontolex*. More recent work on the conversion between TEI Lex-0 and Ontolex in the context of the MorDigital project was presented by Khan et al. [31]. This work is dependent on having quality TEI-encoded data, obtained upstream, which is a fundamental task for the project.

[14]The representation of grammatical and semantic information was omitted to simplify the diagram.

An important component of MorDigital pertains to the modelling of domain ontologies covering the subject fields referred to by the domain labels of the Morais Silva dictionary. While OntoDomLab-Med already covers domains in the medical and health sciences (e.g., medicine, surgery, pharmacy), further work is being carried out in modelling the remaining domains[15]. At the same time, work is being carried out in the SKOS modelling of the MorDigital domain classification, whose first version was already published as linked open data[16]. The latter started from the non-hierarchical list of the 35 domains referred to in the Morais Silva dictionary's list of abbreviations, which will be gradually structured in articulation with the work carried out for the domain ontologies. Further terminological and ontological work will be required to put forward a list of superdomains and hierarchical structure for the domain classification. These are major challenges in the project due to the inherent diversity of specialised fields referred to in the lexicographic articles, including domains absent from the dictionary's list of abbreviations.

Acknowledgments

This paper is supported by (1) the MORDigital – Digitalização do Diccionario da Lingua Portugueza 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

[1] J. P. Silvestre, Bluteau e as origens da lexicografia moderna, INCM, Lisboa, 2009.
[2] T. Verdelho, O dicionário de Morais Silva e o início da lexicografia moderna, in: História da língua e história da gramática: actas do encontro, Universidade do Minho, Braga, 2003, pp. 473–490.
[3] M. Correia, Os dicionários portugueses, Caminho, Lisboa, 2009.
[4] R. Costa, A. Salgado, A. Khan, S. Carvalho, L. Romary, B. Almeida, M. Ramos, M. Khemakhem, T. Tasovac, R. Silva, MORDigital: the advent of a new lexicographical Portuguese project, in: I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek, C. Tiberius (Eds.), Electronic lexicography in the 21st century: post-editing lexicography. Proceedings of eLex 2021, Lexical Computing, Brno, 2021, pp. 312–324.
[5] T. Tasovac, L. Romary, P. Banski, J. Bowers, J. Does, K. Depuydt, T. Erjavec, A. Geyken, A. Herold, V. Hildenbrandt, M. Khemakhem, B. Lehečka, S. Petrović, A. Salgado, A. Witt, TEI Lex-0: a baseline encoding for lexicographic data. Version 0.9.0, 2018. URL: https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

[15] For example, OntoDomLab-Math was recently developed for modelling the superdomain of mathematical sciences: https://github.com/GuidaRamos/OntoDomLab-Math.

[16] Available from: https://vocabs.rossio.fcsh.unl.pt/morais_domains/

[6] L. Romary, TEI and LMF crosswalks, Journal for language technology and computational linguistics 30 (2015) 47–70.
[7] P. Cimiano, J. P. McCrae, P. Buitelaar, Lexicon Model for Ontologies: Community Report, Technical Report, W3C Ontology-Lexicon Community Group, 2016. URL: https://www.w3.org/2016/05/ontolex/.
[8] P. Cimiano, C. Chiarcos, J. P. McCrae, J. Gracia, Linguistic Linked Data: Representation, Generation and Applications, Springer, Berlin, 2020.
[9] J. Bosque-Gil, J. Gracia, J. P. McCrae, P. Cimiano, S. Stolk, F. Khan, K. Depuydt, J. Does, F. Frontini, I. Kernerman, The OntoLex Lemon Lexicography Module: final community report, Technical Report, W3C Ontology-Lexicon Community Group, 2019. URL: https://www.w3.org/2019/09/lexicog/.
[10] F. Khan, A. Bellandi, F. Boschetti, M. Monachini, The Challenges of Converting Legacy Lexical Resources to Linked Open Data using Ontolex-Lemon, in: LDK Workshops, 2017, pp. 1–8. URL: http://ceur-ws.org/Vol-1899/OntoLex_2017_paper_4.pdf.
[11] R. Stanković, R. Stijović, D. Vitas, C. Krstev, O. Sabo, The Dictionary of the Serbian Academy: from the Text to the Lexical Database, in: Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana University Press, Ljubljana, 2018, pp. 941–949. URL: https://dais.sanu.ac.rs/123456789/4927.
[12] J. P. McCrae, C. Tiberius, A. Khan, I. Kernerman, T. Declerck, S. Krek, M. Monachini, S. Ahmadi, The ELEXIS Interface for Interoperable Lexical Resources, in: I. Kosem, T. Zingano Kuhn, M. Correia, J. Ferreira, M. Jansen, M. Pereira, J. Kallas, M. Jakubíček, S. Krek, C. Tiberius (Eds.), Electronic lexicography in the 21st century: proceedings of the eLex 2019 conference, Lexical Computing, Brno, 2019, pp. 642–659. URL: http://hdl.handle.net/10379/15512.
[13] F. Khan, A. Salgado, Modelling Lexicographic Resources using CIDOC-CRM, FRBRoo and Ontolex-Lemon, in: A. Bikakis, R. Ferrario, S. Jean, B. Markhoff, A. Mosca, M. N. Asmundo (Eds.), Proceedings of the International Joint Workshop on Semantic Web and Ontology Design for Cultural Heritage co-located with the Bolzano Summer of Knowledge 2021 (BOSK 2021), CEUR-WS, 2021, pp. 1–12.
[14] TEI Consortium, TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.4.0. Last updated on 19th April 2022, revision ff9cc28b0, 2022. URL: https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html.
[15] R. Costa, S. Carvalho, A. Salgado, A. Simões, T. Tasovac, Ontologie des marques de domaines appliquée aux dictionnaires de langue générale, Langue(s) & Parole 5 (2020) 201–230. URL: https://revistes.uab.cat/languesparole/article/view/v5-costa-etal.
[16] R. Costa, A. Salgado, B. Almeida, SKOS as a key element for linking lexicography to digital humanities, in: K. Golub, Y. H. Liu (Eds.), Information and Knowledge Organisation in Digital Humanities : Global Perspectives, Routledge, London, 2021, pp. 178–204. URL: https://doi.org/10.4324/9781003131816-9.
[17] A. Salgado, R. Costa, T. Tasovac, Applying terminological methods to lexicographic work: terms and their domains, in: A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs, P. Storjohann (Eds.), Dictionaries and Society. Proceedings of the XX EURALEX International Congress, IDS-Verlag, Mannheim, 2022, pp. 181–195.
[18] U. Weinreich, Is a Structural Dialectology Possible?, WORD 10 (1954) 388–400. doi:https:

//www10.1080/00437956.1954.11659535.
[19] S. Landau, Dictionaries: the art and craft of lexicography, 2nd ed ed., Cambridge University Press, Cambridge, 2001.
[20] B. Svensén, A handbook of lexicography: the theory and practice of dictionary-making, Cambridge University Press, Cambridge, 2009.
[21] A. Salgado, R. Costa, T. Tasovac, Improving the Consistency of Usage Labelling in Dictionaries with TEI Lex-0, Lexicography 6 (2019) 133–156. doi:10.1007/s40607-019-00061-x.
[22] M. O. Ptaszyński, Theoretical Considerations for the Improvement of Usage Labelling in Dictionaries: A Combined Formal-Functional Approach, International Journal of Lexicography 23 (2010) 411–442. doi:10.1093/ijl/ecq029.
[23] M. Vrbinc, A. Vrbinc, Diasystematic Information in the "Big Five": A Comparison of Print Dictionaries, CD-ROMS/ DVD-ROMS and Online Dictionaries, Lexikos 25 (2015) 424–445. doi:10.5788/25-1-1306.
[24] F. J. Hausmann, Die Markierung in einem allgemeinen einsprachigen Wörterbuch: eine Übersicht, in: F. J. Hausmann, O. Reichmann, H. E. Wiegand, L. Zgusta (Eds.), Wörterbücher. Ein internationales Handbuch zur Lexikographie. Erster Teilband, Walter de Gruyter, Berlin, 1989, pp. 649–657.
[25] H. Bergenholtz, S. Tarp, Manual of Specialised Lexicography: The Preparation of Specialised Dictionaries, John Benjamins, Amsterdam, 1995.
[26] J. Milroy, L. Milroy, Authority in Language: Investigating Standard English, Routledge, London, 1990.
[27] H. Jackson, Lexicography: An Introduction, Routledge, London, 2002.
[28] B. Atkins, M. Rundell, The Oxford Guide to Practical Lexicography, Oxford University Press, New York, 2008.
[29] ISO 24613-1, Language resource management - Lexical markup framework (LMF) - Part 1: Core model, ISO, Geneva, 2019.
[30] A. M. 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, Officina de Simão Thaddeo Ferreira, Lisboa, 1789. URL: https://purl.pt/29264.
[31] F. Khan, A. Salgado, R. Costa, S. Carvalho, L. Romary, B. Almeida, M. Khemakhem, R. Silva, T. Tasovac, Interlinking lexicographic data in the MORDigital project, Mykolas Romeris University, Vilnius, 2022.