Guidelines for Encoding Domain Labels for Linked Data Lexical Resources in RDF

Fahad Khan, Ana Salgado, Rute Costa, Margarida Ramos, Sara Carvalho, Raquel Silva and Bruno Almeida

Domain Labels – An Introduction

In the context of lexicography, the term *domain label* is commonly used to denote a usage label[1] assigned to a sense and serving as a 'marker which identifies the specialised field of knowledge in which a lexical unit is mainly used' (Salgado, Costa & Tasovac, 2019)[2]. These labels are used 'para señalar el léxico temáticamente especializado, en contraposición al léxico común' [to signal the thematically specialised lexicon in contrast to the common lexicon] (Estopà, 1998, p. 1) and are generally expressed in the form of abbreviations representing individual domains, especially in the case of paper dictionaries[3]. Throughout this work, we will use the term domain label to refer both to the abbreviations observed in individual dictionary entries (e.g., *Geograf.*) as well as to fuller

[1] Although we do not go into any detail on the broader topic of usage labels here, it is important to understand that the association of such labels with a lexical unit implies that the latter moves away 'in a certain respect, from the main bulk of items described in a dictionary, and that its use is subject to some kind of restriction' (Svensén, 2009, p. 313). The need to label certain deviations and restrictions in the use of a term (such as, for instance, when it is associated with a familiar register or if it belongs to a specialised domain) originated in what is currently called *marking* or *diasystematic marking* (Hausmann, 1989, p. 651).

[2] Note that the designation 'domain label' is not universally accepted. Atkins and Rundell (2008), referring to 'linguistic labels', classified specialised vocabulary as 'domains' (p. 182); they are termed 'field labels' by Verkuyl, Janssen and Jansen (2003, p. 7), 'marcas técnicas' by Fajardo (1996/1997), 'marca de materia' (Martínez de Sousa, 1995), 'marca terminológica' in Lara (1997), 'marcas temáticas' in Estopà (1998), 'field label' (Hartmann & James, 1998/2002), 'marca de especialidad' (Nomdedeu Rull, 2008), or 'diatechnical information/marking' (Hausmann, 1989; Svensén, 2009). We settled on the term 'domain label' as we felt it was both transparent and recognisable to lexicographers as well as serving as a beacon for terminologists.

[3] We define a domain as a 'field of special knowledge' (ISO 1087, 2019, p. 1): this definition has the advantage of being both transparent and sufficiently comprehensive. Taking the complexity of domain knowledge into consideration, Sager (1990) states that '[i]n 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).

versions of these abbreviations often found in the front matter of a dictionary, e.g. GEOGRAFIA [GEOGRAPHY]).

Although domain labels are commonly associated with individual lexical unit senses, they can also be assigned to individual entries (this is very useful in case a lexical unit is only associated with a specialised sense, which effectively make this unit a term in itself) as well as other components of an entry. Domain labels can, moreover, be organised in taxonomies or thesauri, which can help make lexicons easier to navigate and to query. Although such labels play an essential role in lexical resources, and especially in lexicographic resources, so far there has not been much work on modelling them in linked data lexicons in a way that better exploits the possibilities of the Semantic Web stack (see however Almeida et al., 2022).

To help correct this state affairs, we will present a series of guidelines for encoding domain label information in RDF using three linked data vocabularies, namely OntoLex-Lemon, SKOS, and lexicog. These guidelines will be illustrated by a series of examples from two Portuguese language dictionaries, one contemporary and the other historical. Namely, we will take our examples from the *Dicionário da Língua Portuguesa* or DLP-ACL (ACL, 2023), a recent digital version of the Academia das Ciências Portuguese language dictionary, and the 19th century *Diccionario da Lingua Portugueza de António de Morais Silva* which is currently being published as a digital edition as part of the Portuguese national project MorDigital.

Requirements

In the rest of the document, we will assume a basic familiarity with the OntoLex-Lemon vocabulary, the Lexinfo vocabulary as well as the SKOS vocabulary.

Best Practises for Encoding Domain Labels

The original predecessor of OntoLex-Lemon, namely, the LExicon Model for ONtologies (*lemon*), allowed for the addition of topic information to entries via the use of the `lemon:topic` property, along with `lemon:context` to specifying the technical register of a given sense. While OntoLex-Lemon did not retain these properties, the OntoLex-Lemon guidelines instead suggest the use of the `dct:subject` property to specify: > under which conditions (context, register, domain, etc) it is valid to regard the lexical entry as having the ontological entity as meaning.

The same guidelines also recommend the use of the `ontolex:usage` property which is defined as specifying the > usage conditions or pragmatic implications when using the lexical entry to refer to the given ontological meaning

This property has the domain of `ontolex:LexicalSense` and the range `rdfs:Resource`. Moreover, the lexinfo vocabulary[4], defines a series of sub-

[4] Here and throughout this document when we mention lexinfo we are referring to lexinfo

properties of `ontolex:usage` including `lexinfo:domain` which is defined as a: >usage marker which identifies the specialized field of knowledge in which a lexical unit is mainly used.

Ontolex therefore offers us a way of marking a lexical entry as belonging to a certain domain and a way of specifying that a specific sense of an entry is associated with a particular domain. When it comes to encoding the domain label itself, we suggest encoding it as a instance of the SKOS class `Concept` and using the `skos:narrower` and `skos:broader` relations to encode the relations between different domains. We therefore suggest the following steps when encoding domain label information in linked data lexical resources.

---

1. Domain labels should be encoded as individuals of the class `skos:Concept`. Hierarchical relationships between individual domain labels should be encoded using the `skos:narrower` and `skos:broader` properties. In the case of retrodigitised and non-native-born dictionaries, it may be that the same domain label is not consistently encoded using the same string; in such situations, we recommend using `skos:preflabel` and `skos:altlabel` to list the different versions of the same label (with the former being used to encode the version(s) found in the front matter and the latter its variants).
2. In case the whole entry is marked as (or interpreted by the encoder as) belonging to a given domain we recommend encoding this information using `dcterms:subject` with the entry as subject and the relevant domain label (encoded as `skos:Concept`, see above) as object.

3. In case a single sense is marked as (or is interpreted by the encoder as) belonging to a domain, we recommend using `lexinfo:domain` with the entry as subject and the relevant `skos:Concept` as object.
4. In other cases where any other part of the entry is marked with a domain label, once again we recommend the use of `dcterms:subject`.

Examples

Namespaces

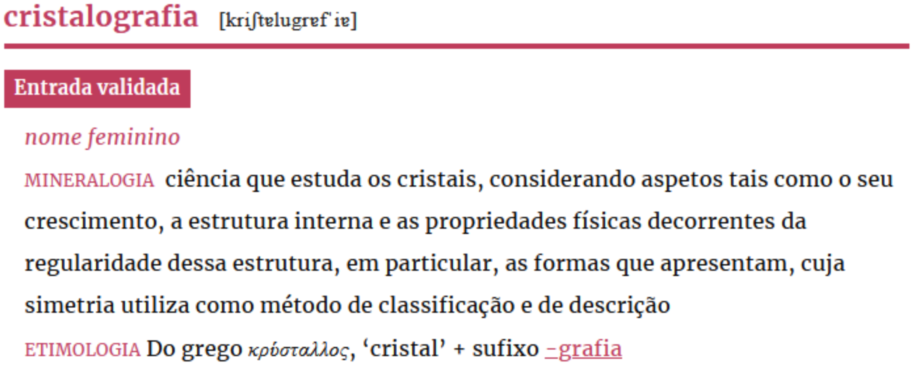
In the examples that follow, we use the following namespaces:

```
@prefix lexinfo: <http://www.lexinfo.net/ontology/3.0/lexinfo#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix lexicog: <http://www.w3.org/ns/lemon/lexicog#> .
```

3.0.

Encoding hierarchical domain labels in the *Academia* dictionary

In the first example, we show how to encode an entry which has a unique sense that has been marked with a domain label and where the domain referred to is part of a hierarchy of domains. The entry in question is for the Portuguese lexical unit *cristalografia* ‘crystallography’ and comes from the *Academia* dictionary. As the following figure shows, this entry has one sense which is marked with the label MINERALOGIA referring to the domain of mineralogy.



cristalografia [kriʃtɐlugrɐf'iɐ]

Entrada validada

nome feminino

MINERALOGIA ciência que estuda os cristais, considerando aspetos tais como o seu crescimento, a estrutura interna e as propriedades físicas decorrentes da regularidade dessa estrutura, em particular, as formas que apresentam, cuja simetria utiliza como método de classificação e de descrição

ETIMOLOGIA Do grego *κρύσταλλος*, ‘cristal’ + sufixo -grafia

Some additional information relevant to this example is that the domain of MINERALOGIA is a subdomain of the GEOLOGIA ‘geology’ domain in the *Academia* dictionary subject hierarchy that belongs to CIÊNCIAS DA TERRA ‘earth sciences’ superdomain.

We can represent these domains and their interrelations as follows using the SKOS vocabulary:

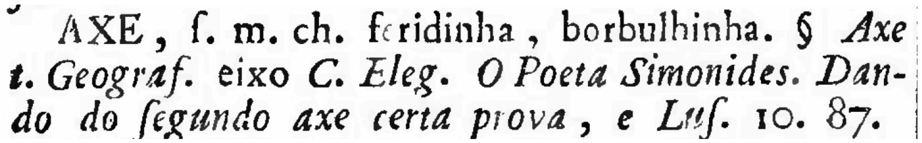
```
<http://example.org/class/mineralogia> rdf:type  skos:Concept;
  skos:prefLabel "mineralogia"@pt;
  skos:prefLabel "minerology"@en;
  skos:narrower <http://example.org/class/geologia> .
<http://example.org/class/geologia> rdf:type  skos:Concept;
    skos:prefLabel "geologia"@pt;
    skos:prefLabel "geology"@en;
    skos:narrower <http://example.org/class/ciencas_da_terra> ;
    skos:broader <http://example.org/class/mineralogia> .
<http://example.org/class/ciencias_da_terra> rdf:type  skos:Concept;
    skos:prefLabel "ciencias da terra"@pt;
    skos:prefLabel "earth sciences"@en;
    skos:broader <http://example.org/class/mineralogia> .
```

In the entry itself, we link the (single) sense of the entry for *cristalografia* (note that the sense is a blank node in the current example) to the domain <http://example.org/class/mineralogia> via the lexinfo:domain property.

```
  <http://example.org/class/DLP_cristalografia> a ontolex:LexicalEntry ;
   lexinfo:etymology [ rdf:value "Do grego       cristal + sufixo -grafia"@pt ] ;
   lexinfo:gender lexinfo:feminine ;
   lexinfo:partOfSpeech lexinfo:noun ;
    ontolex:canonicalForm [
                        ontolex:phoneticRep "kri t lu r fi "@pt ;
                        ontolex:writtenRep "cristalografia"@pt
                        ] ;
     ontolex:sense [ lexinfo:domain <http://example.org/class/mineralogia>;
                       skos:definition
            """ciência que estuda os cristais,
            considerando aspetos tais como o seu crescimento,
            a estrutura interna e
            as propriedades físicas decorrentes da regularidade dessa estrutura,
                        em particular, as formas que apresentam,
            cuja simetria utiliza como método de
                        classificação e de descrição"""@pt ] .
```

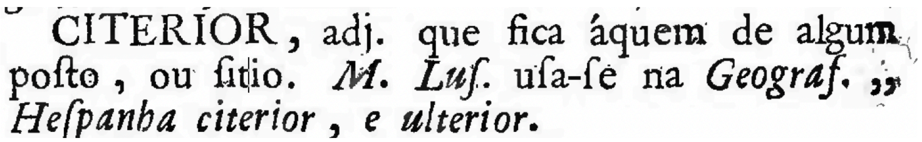
Encoding hierarchical domain labels in the *Morais* dictionary

Our second example is from the encoding of a retrodigitised dictionary, the *Diccionario da Lingua Portugueza de António de Morais Silva.* In this example we will see the use of variants for the same domain label (different abbreviations, italics, bold, formulae in the definitions that point to a domain, etc.). We will look at two individual entries in what follows. The first is the entry for the polysemic word *axe* 'pimple, axle' and the second is the entry for *citerior* 'on the near side of something'. Both are shown in the figures below.



AXE, ſ. m. ch. feridinha, borbulhinha. § *Axe t. Geograf.* eixo *C. Eleg. O Poeta Simonides. Dando do ſegundo axe certa prova, e Luſ.* 10. 87.

Figure 1: Axe Example



CITERIOR, adj. que fica áquem de algum poſto, ou ſitio. *M. Luſ.* uſa-ſe na *Geograf.* „ *Heſpanha citerior, e ulterior.*

Figure 2: Citerior Example

Both of these entries include a domain label pertaining to the domain of GE-

OGRAPHY. In the first entry, this is referred to as "t. Geograf."; in the second example "*Geograf.*". We encode this marker as follows:

```
 <http://example.org/class/geografia> rdf:type  skos:Concept ;
     skos:prefLabel "t. Geograf."@pt;
     skos:altLabel "Geograf."@pt.
```

Note here the two different labels for the domain, with '*t.Geograf*' as the preferred label (since it is listed in the dictionary's front matter).

Moving onto the entry for *axe*, we can encode it as follows:

```
 <http://example.org/individual/MORAIS.1.DLP.AXE> a ontolex:LexicalEntry ;
     lexinfo:gender lexinfo:masculine ;
     lexinfo:partOfSpeech lexinfo:noun ;
     ontolex:canonicalForm [ ontolex:writtenRep "AXE"@pt ] ;
     ontolex:sense [
                     lexinfo:socioCultural [ rdf:value "ch." ] ;
                     skos:definition "feridinha, borbulhinha"@pt
                    ],
                    [
                     lexinfo:domain <http://example.org/class/geografia> ;
                     skos:definition "eixo"@pt ;
                     lexicog:usageExample [
            dcterms:source "C. Eleg. O Poeta Simonides . ";
                            rdf:value "Dando do segundo axe certa prova"@pt ];
                     lexicog:usageExample [
            dcterms:source "Luſ. . 10. 87. . "]
                     ] .
```

Note that the entry has two different senses (both of these represented as blank nodes)[5]. The second sense is the relevant one in our case; note also the two usage examples associated with the sense. Once again we use the `lexinfo:domain`[6].

```
<http://example.org/instance/MORAIS.1.DLP.CITERIOR> a ontolex:LexicalEntry ;
    ontolex:canonicalForm [ ontolex:writtenRep "CITERIOR"@pt ] ;
    lexinfo:partOfSpeech lexinfo:adjective ;
    ontolex:sense [
                skos:definition "que fica áquem de algum poſto, ou ſitio"@pt ;
                lexicog:usageExample [
                            dcterms:source
            " M. Luſ. ,, usa-se na t. Geograf. Hespanha citerior, e ulterior . ";
                            dcterms:subject  <http://example.org/class/geografia>] ] .
```

[5] We can order these two senses using the `lexicog:LexicographicComponent`class, see the lexicog guidelines. We decided not to do this in the current case in the interests of keeping the exposition as simple as possible.

[6] In order to keep the example simple we haven't added any structured bibliographic information, even though this can be easily done using a number of linked data vocabularies such as [...].

Note that in this case we associate the domain label with the usage example rather than the entry or even the sense, making use, in this case of `dcterms:subject`.

Acknowledgements

This work presented here was supported by a short term scientific mobility grant from the COST Action NexusLinguarum (CA18209) supported by COST (European Cooperation in Science and Technology). It was also supported by the MORDigital – Digitalização do Diccionario da Lingua Portugueza de António de Morais Silva [PTDC/LLT-LIN/6841/2020] project financed by Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia.

References

Svensén, B. (2009). A Handbook of Lexicography: The Theory and Practice of Dictionary Making. Cambridge: Cambridge University Press. Almeida, B., Costa, R., Salgado, A., Ramos, M., Romary, L., Khan, F., … & Tasovac, T. (2022). Modelling Usage Information in a Legacy Dictionary: From TEI Lex-0 to Ontolex-Lemon.

Hausmann, F. J. (1989). Die Markierung in eineim allgemeinen einsprachigen Wörterbuch: eine Übersicht. In F. J. Hausmann, O. Reichmann, H. E. Wiegand, L. Zgusta (Eds.), Wörterbücher. Ein internationales Handbuch zur Lexikographie (pp. 649–657). Berlin: Walter de Gruyter.

Salgado, A., Costa, R. & Tasovac, T. (2019). Improving the consistency of usage labelling in dictionaries with TEI Lex-0. Lexicography: Journal of ASIALEX, 6(2), 133–156. doi:10.1007/s40607-019-00061-x.

Estopà, R. B. (1998). El léxico especializado en los diccionarios de lengua general: las marcas temáticas. Revista de la Sociedad Española de Linguística, 28(2), 359–387.

Atkins, B. T. S., & Rundell, M. (2008). The Oxford Guide to Practical Lexicography. New York: Oxford University Press.

Verkuyl, H. J., Janssen, M., & Jansen, F. (2003). The codification of usage by labels. In Sterkenburg, P. (Ed.), A practical guide to lexicography (pp. 297–311). Amsterdam: John Benjamins. doi:10.1075/tlrp.6.33ver.

Sager, J. C. (1990). A practical course in terminology processing. Amsterdam: John Benjamins Publishing Company.

ISO 1087. (2019). Terminology Work – Vocabulary – Part 1: Theory and Application. Geneva: International Organization for Standardization.

Almeida, B., Costa, R., Salgado, A., Ramos, M., Romary, L., Khan, F., Carvalho, S., Khemakhem, M., Silva, R., & Tasovac, T. (2022). Modelling Usage Information in a Legacy Dictionary: From TEI Lex-0 to Ontolex-Lemon.

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.

ACL (2023). Dicionário da Língua Portuguesa. Salgado, A. (Coord.). Lisboa: Academia das Ciências de Lisboa. https://dicionario.acad-ciencias.pt/