MorDigital

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 and Toma Tasovac

Workshop on Computational Methods in the Humanities 2022 (COMHUM 2022)

10 June 2022

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

NOVAFCSH
FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS
UNIVERSIDADE NOVA DE LISBOA

FCT Fundação para a Ciência e a Tecnologia

Summary

1. The MorDigital project: digitisation of a legacy dictionary
2. Usage information in lexicographic resources
3. TEI Lex-0, encoding a lexicographic article
4. Ontolex-Lemon, modelling a lexicographic article
5. Final remarks

The MorDigital project: digitisation of a legacy dictionary

- The MorDigital project aims **to digitise the first three editions of *Diccionario da Lingua Portugueza*** by António de Morais Silva (1789; 1813; 1823), commonly referred to as “the Morais dictionary” (Costa et al., 2021).
- MorDigital embodies a paradigm in which lexicography converges with several domains, such as terminology, ontologies, linked data and digital humanities.
- The digital versions will be structured based on standards for encoding and modelling lexical resources, namely TEI Lex-0, Lexical Markup Framework (LMF) and Ontolex-Lemon.
- The project will result in the digital preservation of the Morais dictionary through an open access platform, where its contents will be fully searchable and accessible to the community.
- 36 months grant, 2021-2023 (FCT – Fundação para a Ciência e Tecnologia).

The MorDigital project: digitisation of a legacy dictionary

- International and multidisciplinary team, with participants from the following institutions:
  - NOVA CLUNL – Centro de Linguística da Universidade NOVA de Lisboa, Portugal.
  - Academia das Ciências de Lisboa, Portugal.
  - Istituto Di Linguistica Computazionale 'A. Zampolli', Italy.
  - CLLC – Centro de Línguas, Literaturas e Culturas da Universidade de Aveiro, Portugal.
  - Inria, ALMAnaCH team, France.
  - ArcaScience, France.
  - BCDH – Belgrade Centre for Digital Humanities, Serbia.

The MorDigital project: digitisation of a legacy dictionary

- Analysis of macro- and microstructural components.
- Typologies of lexicographic information.

- Re-OCRisation of public domain versions of the Morais dictionary.

- Generation of a TEI Lex-0 and LMF compliant encoding using GROBID-Dictionaries, a machine learning library based on cascading CRF models.

- Construction and publication of the Morais Domain Classification.
- Selection of ontological predicates for modelling dictionary data (from Ontolex, LexInfo, etc.).

- RDF conversion using tei2ontolex, a XSL-based tool.
- Alignment (entries, senses).

- Browsing, simple and advanced search functionalities.
- REST API for exposing dictionary data.

The MorDigital project: digitisation of a legacy dictionary

https://mordigital.fcsh.unl.pt/

Diccionario da Lingua Portugueza by Morais Silva

- As the **first Portuguese monolingual dictionary**, it was instrumental in normalising this language and was the model for subsequent Portuguese dictionaries.
- The Morais dictionary was devised in the **Age of Enlightenment**, and was influenced by other modern language dictionaries published in Europe in the 16th and 17th centuries.
- In the 1st ed., authorship is attributed to **Rafael Bluteau**, a Portuguese priest and lexicographer, whose Portuguese-Latin Vocabulary (10 vols., 1712-1728) was the basis for the Morais dictionary.
- Morais directly oversaw the 2nd (1813) and 3rd (1823) editions, which greatly overhauled the dictionary.

Frontispiece of Morais (1789, 1813, 1823)

Two volumes
A to K, 752 pp.
L to Z, 541 pp.

Two volumes
A to E, 889 pp.
F to Z, 886 pp.

Two volumes
A to F, 962 pp.
G to Z, 872 pp.

Usage information in lexicographic resources

- Usage, or diasystematic, information represents **constraints** on the use of lexical items to certain contexts, or to a subset of language users.
- Dictionaries traditionally include usage information in the lexicographic articles as labels, notes, or within the definitions themselves.
- There are several typologies of usage information, mostly adapted from Hausmann (1989), which organise it according to several criteria (e.g. time, place, degree of specialty).
- In the TEI guidelines, the `<usg>` element is put forward for marking usage information in dictionaries.
  - It is a typed element in TEI Lex-0, with `@type` as a mandatory attribute.
  - It uses a specified list of values (Salgado et al., 2019), with `"hint"` as the default value.

Typologies of usage information

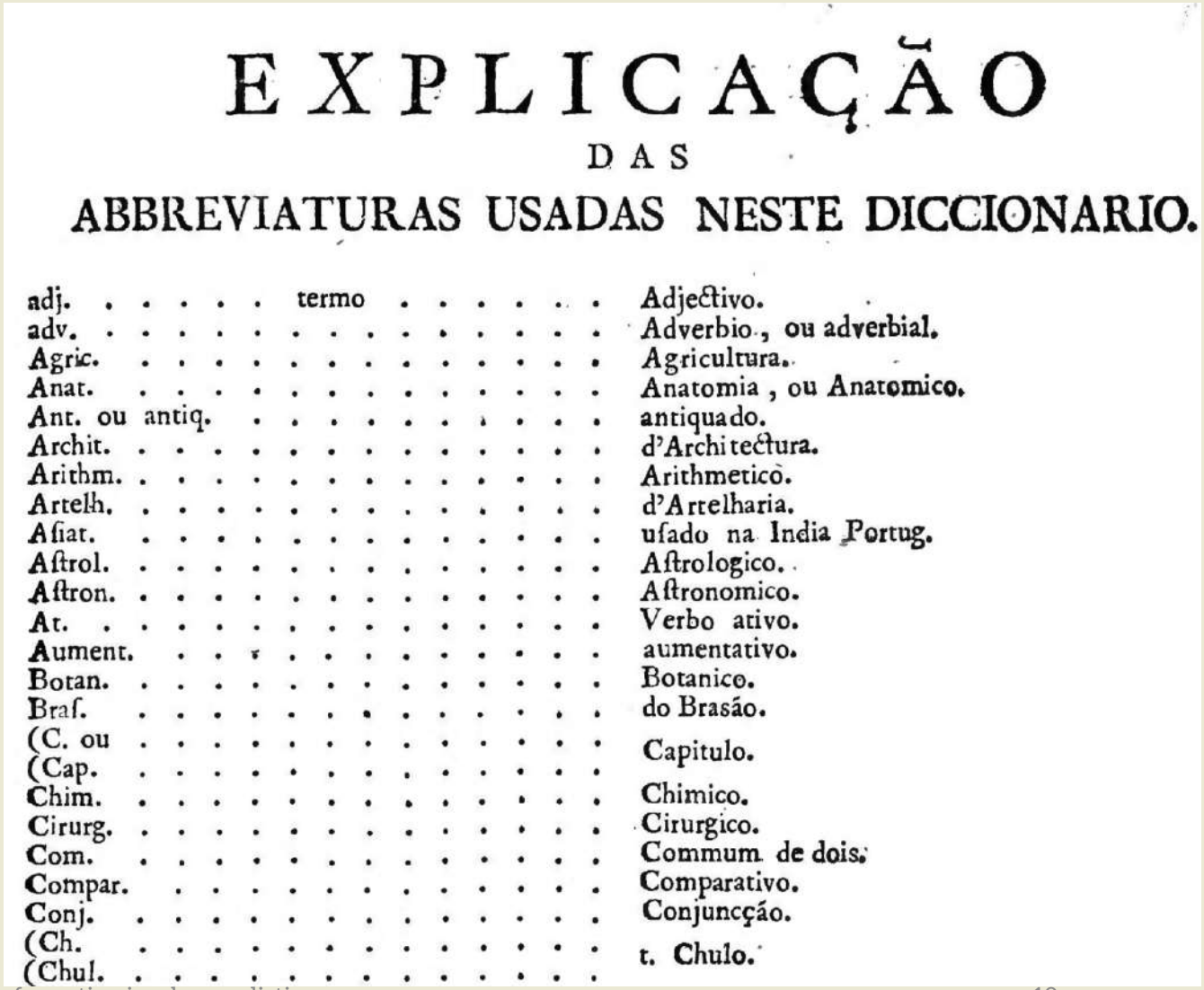
Adapted from: Salgado, Costa & Tasovac (2019)

MorDigital

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/ geographic 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	SOCIO-CULTURAL	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. (=figurative), lit. (= literal)

Usage information in the Morais dictionary

- The Morais dictionary includes labels associated to senses, usually through **abbreviations** typeset in italics.
- Some definitions also include usage information, which is given in a more verbose form.
- The analysis of the lists of abbreviations allowed to put forward a typology of usage information in the Morais dictionary.



EXPLICAÇÃO
DAS
ABBREVIATURAS USADAS NESTE DICCIONARIO.

adj. termo	Adjectivo.
adv.	Adverbio, ou adverbial.
Agric.	Agricultura.
Anat.	Anatomia, ou Anatomico.
Ant. ou antiq.	antiquado.
Archit.	d'Architectura.
Arithm.	Arithmetico.
Artelh.	d'Artelharia.
Aſiat.	uſado na India Portug.
Aſtrol.	Aſtrologico.
Aſtron.	Aſtronomico.
At.	Verbo ativo.
Aument.	aumentativo.
Botan.	Botanico.
Braſ.	do Brasão.
(C. ou (Cap.	Capitulo.
Chim.	Chimico.
Cirurg.	Cirurgico.
Com.	Commum de dois.
Compar.	Comparativo.
Conj.	Conjuncção.
(Ch. (Chul.	t. Chulo.

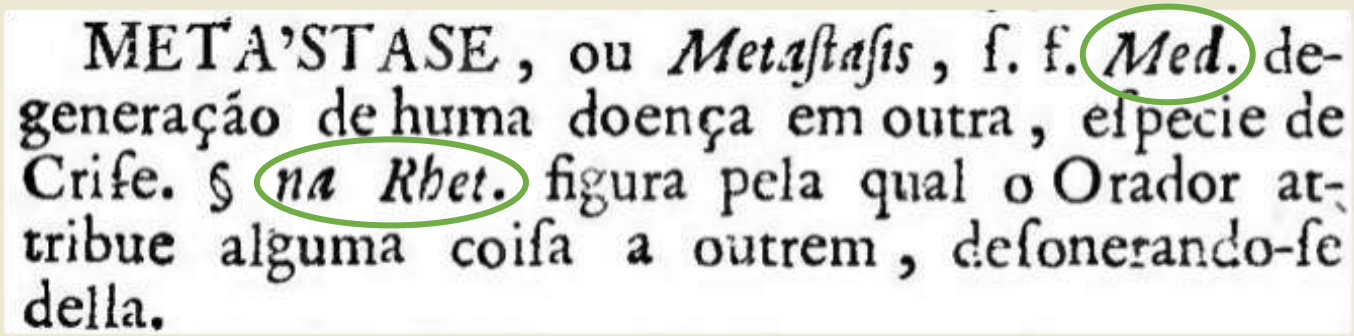
Usage information in the Morais dictionary

- **Diatechnical information (`domain`)**: usage restricted to a subject domain (e.g. *Med.*)
- **Diatextual information (`textType`)**: usage restricted to a text or discourse type (e.g. *Poet.*)
- **Diaevaluative information (`attitude`)**: usage indicates the speaker's attitude (e.g. *t. chulo*)
- **Diastratic information (`socioCultural`)**: usage restricted to a social group (e.g. *Vulg.*)
- **Diaphasic information (`socioCultural`)**: usage restricted to a language register (e.g. *Fam*.)

Usage information in the Morais dictionary

- **Diatopic information (`geographic`)**: usage restricted to a regional variety (e.g. *Asiat.*)
- **Diachronic information (`temporal`)**: usage associated with a period in history (e.g. *Ant.*)
- **Diaintegrative information (`hint`)**: indicating a loanword (e.g. *Lat.*)
- **Diafrequential information (`frequency`)**: indicating frequency of occurrence (e.g. *P. us.*)

Usage information: a few examples (1st ed., 1789)


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.

The different senses of the lexicographic article are constrained to individual subject fields (medicine and rhetoric).

METASTASIS , n. f. Med. degeneration of a disease into another , a kind of Crisis. § in Rhet. figure by which the Speaker attributes something to another , disowning it.

[Freely translated into English]

Usage information: a few examples (1st ed., 1789)

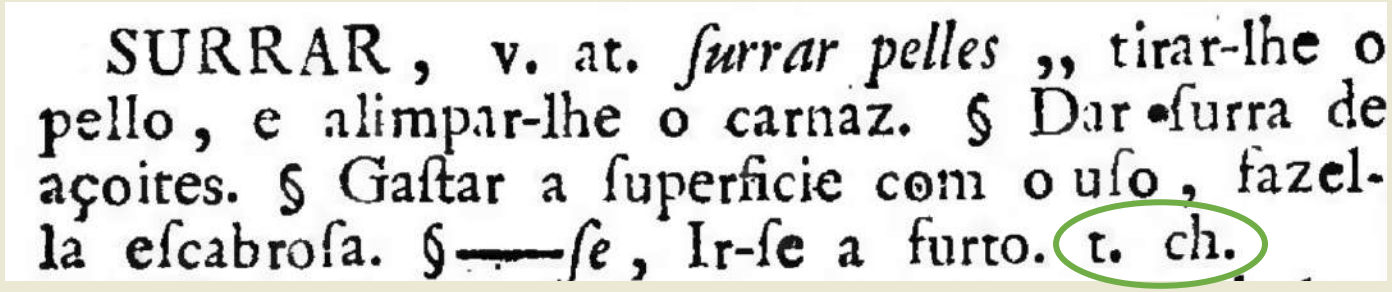

N'ELLE, ſ. m. arroz com caſca, na Aſia.

The usage of the lexical item is restricted to speakers from a given geographic location (Portuguese India).

NELLE , n. m. rice with husk , in Asia.

[Freely translated into English]

Usage information: a few examples (1st ed., 1789)



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

The pronominal usage of the verb is marked as being ironic or malicious, meaning “to run away”, “to remove oneself”

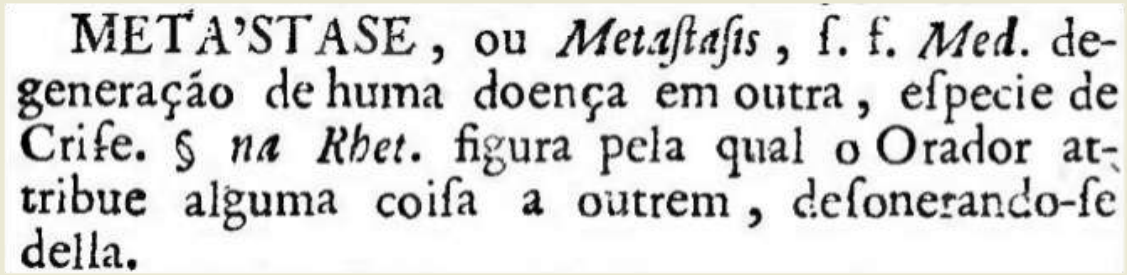
FLESH , v. t. to flesh hides ,, to remove its hair, and clean the split side. § To whip. § To worn out the surface with use , making it rough. § [Pronominal usage] , to run away. Ironic/malicious.

[Freely translated into English]

TEI Lex-0: a baseline encoding for lexicographic data

- TEI Lex-0 is rooted in the Guidelines of the Text Encoding Initiative, and is delivered as a customisation of the TEI schema (Tasovac et al., 2018).
- It aims to facilitate the interoperability of lexicographic resources.
- Developed in the framework of the Lexical Resources WG of DARIAH-EU, and ELEXIS-EU.
- There has been a convergence between TEI Lex-0 and LMF in the past few years (Romary, 2015).

Encoding an article in TEI Lex-0



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.

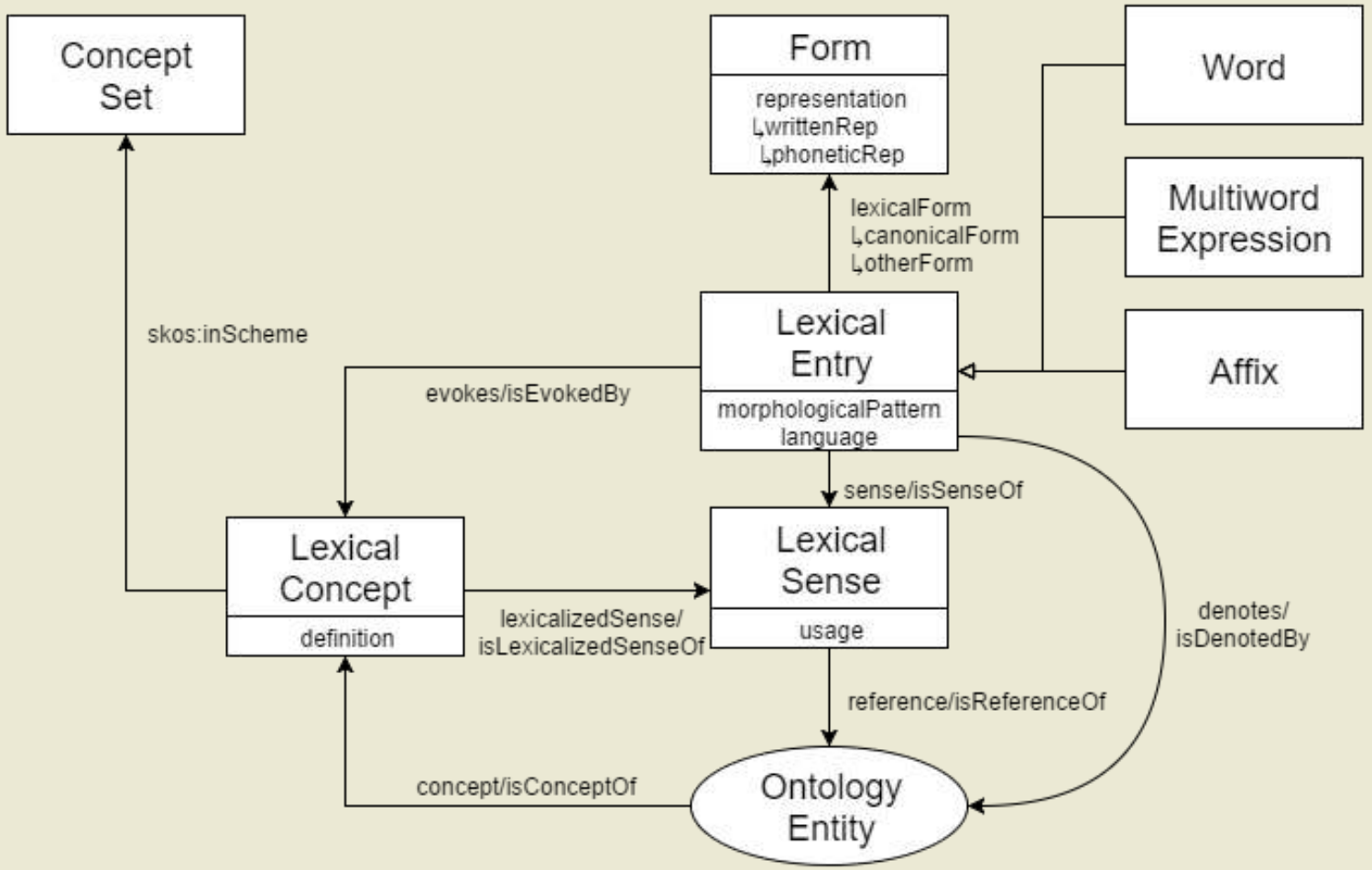
- Usage information is encoded through the `<usg>` element within each sense.
- The information is typed as **domain labels**, identifying the subject field of each of the words' senses.

```
<entry xmlns="http://www.tei-c.org/ns/1.0"
type="monolexicalUnit" xml:lang="pt"
xml:id="MORAIS_1.metastase">
    <form type="lemma">
        <orth>METÁSTASE</orth>
        <pc>, ou</pc>
        <form type="variant">
            <orth>Metastasis</orth>
        </form>
        <pc>,</pc>
        <gramGrp>
            <gram type="pos"
norm="NOUN">s.</gram>
            <gram type="gen">f.</gram>
        </gramGrp>
        <sense xml:id="MORAIS_1.metastase_1">
        <usg type="domain">Med.</usg>
        <def>degeneração de huma doença em
outra, espécie de Crise</def>
        </sense>
    <sense xml:id="MORAIS_1.metastase_2">
        <pc>na</pc>
        <usg type="domain">Rhet.</usg>
        <def>figura pela qual o Orador attribue
alguma coisa a outrem , desonerando-se
della.</def>
</entry>
```

Ontolex-Lemon

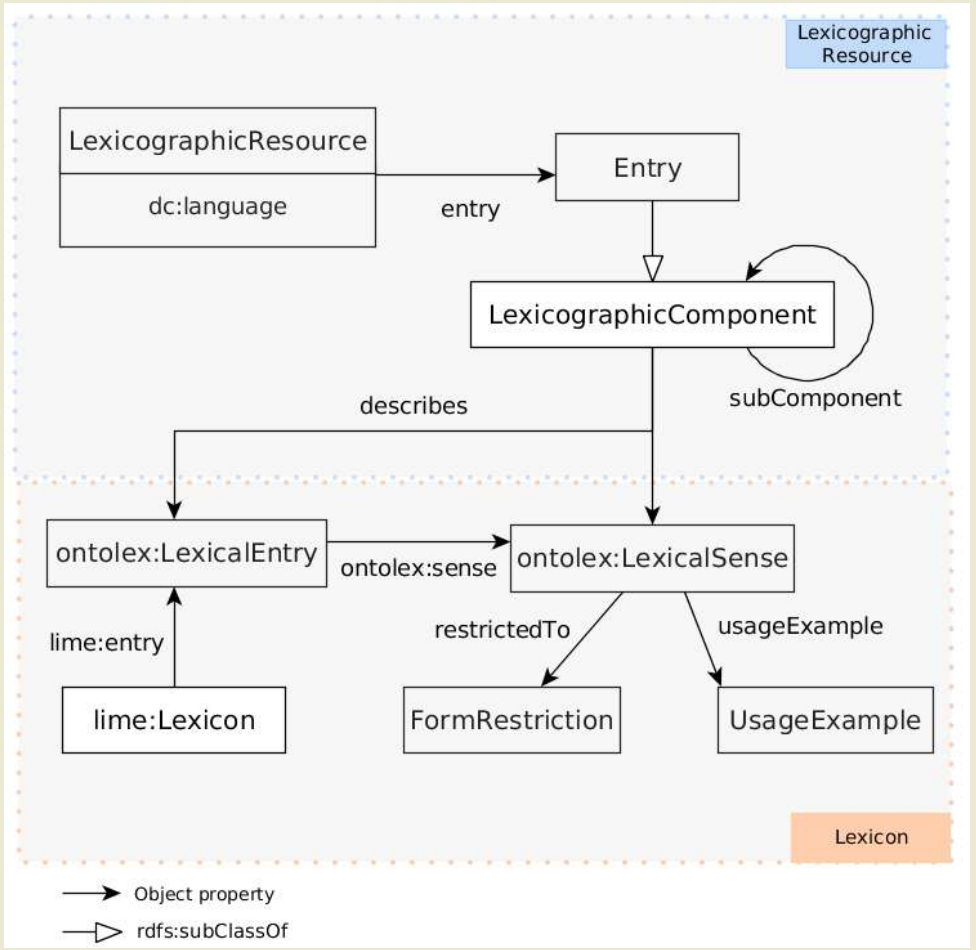
- Ontolex-Lemon is a model originally developed by a W3C community group for enriching ontologies with lexical information (Cimiano, McCrae & Buitelaar, 2016).
- It has since become a *de facto* standard for publishing lexical resources as linked data (Cimiano et al., 2020).
- The **core module** includes classes and properties for modelling information associated with lexical entries, (e.g., forms, senses), and linking them to ontology elements.
- The **lexicography module**, developed more recently, includes additional elements for modelling dictionaries, whose entries may describe items from different POS, or different forms belonging to a lexical entry (e.g., the past participle).
- Usage information can be modelled by reusing elements from **lexinfo**, an ontology of linguistic data categories (https://lexinfo.net/).

Ontolex-Lemon: core module



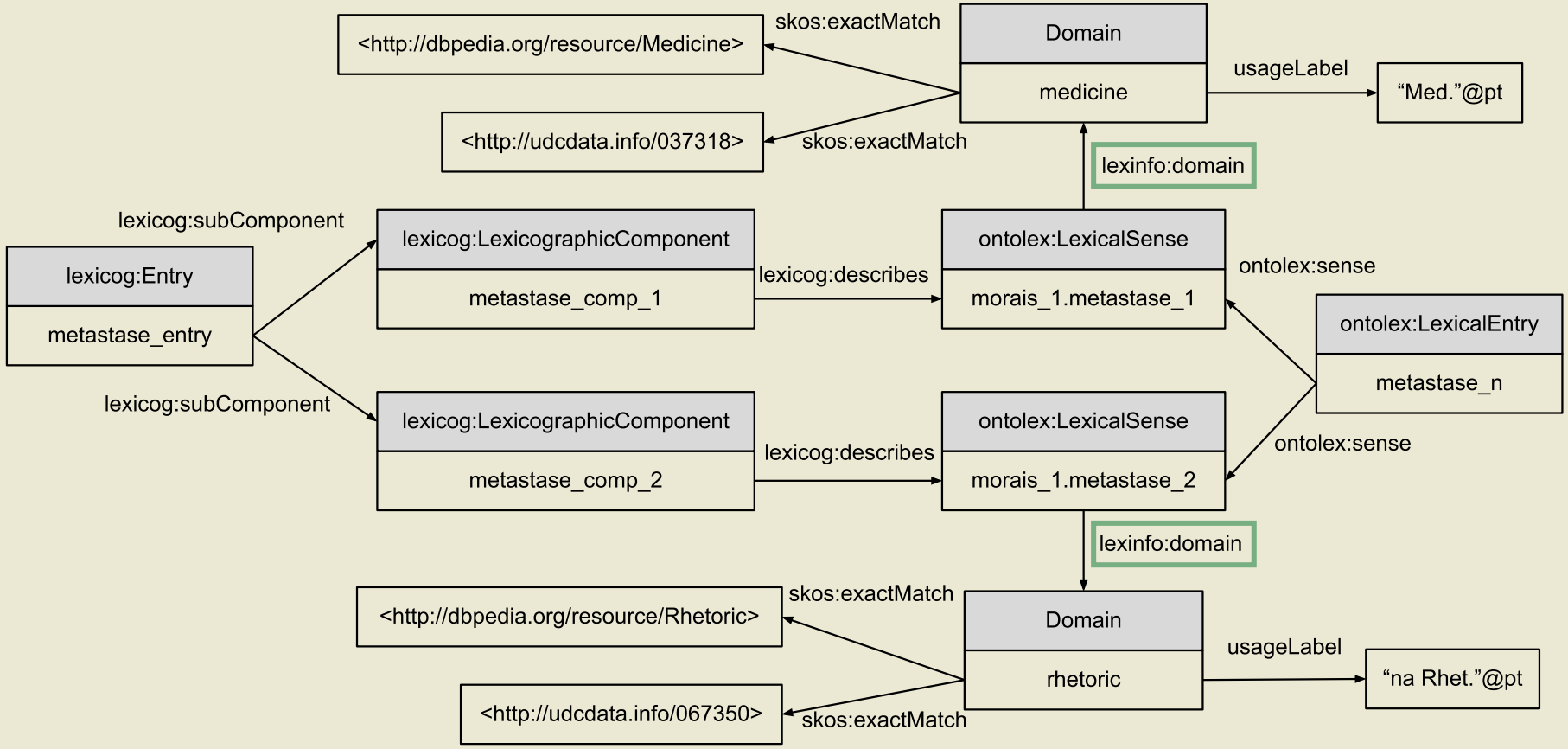
Source: https://www.w3.org/2016/05/ontolex/

Ontolex-Lemon: lexicography module



Source: https://www.w3.org/2019/09/lexicog/

Modelling an article in Ontolex-Lemon



- Lexicog allows to distinguish the **lexical level** from its **lexicographic description** in the Morais dictionary.
- `lexinfo:domain` is used for indicating the subject fields, while the actual labels (as they appear in Morais) are indicated through a custom property.
- A **domain classification** is currently being modelled in SKOS, containing relevant information about the domains in Morais (e.g., usage labels, definitions, alignment with knowledge bases)

Final remarks

- The MorDigital project aims to preserve and provide access to the contents of the first Portuguese monolingual dictionary, which will be a significant contribution to the community.
- The project draws contributions from multiple domains, from lexicography to terminology, ontologies, linked data and digital humanities, embodying a new paradigm for lexicography.
- The methodologies and resources developed for MorDigital should be reusable in other projects involving the retrodigitisation of legacy dictionaries.

Final remarks

- The Morais dictionary includes diverse usage information, usually represented through abbreviated labels, but sometimes included within definitions, which could pose a challenge for machine-learning models.
- TEI Lex-0 will enable the encoding of the Morais dictionary in an interoperable way, providing the basis for visualising, querying and mining dictionary data in the MorDigital platform.
- The linked data conversion of the Morais dictionary enables a further way to explore and enrich its data, and will constitute an important contribution to the Portuguese section of the linguistic linked open data cloud.

Acknowledgements

- 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
- 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
- European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS) (European Lexicographic Infrastructure).

MorDigital

Thank you!

Bruno Almeida: brunoalmeida@fcsh.unl.pt

Rute Costa: rute.costa@fcsh.unl.pt

Ana Salgado: anasalgado@fcsh.unl.pt

Margarida Ramos: mvramos@fcsh.unl.pt

Laurent Romary: laurent.romary@inria.fr

Anas Fahad Khan: fahad.khan@ilc.cnr.it

Sara Carvalho: sara.carvalho@ua.pt

Mohamed Khemakhem: medkhemakhemfsegs@gmail.com

Raquel Silva: raq.silva@fcsh.unl.pt

Toma Tasovac: ttasovac@humanistika.org

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

NOVAFCSH
FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS
UNIVERSIDADE NOVA DE LISBOA

FCT Fundação para a Ciência e a Tecnologia