MorDigital

Interlinking Lexicographic Data in the MORDigital Project

Anas Fahad Khan, CNR-ILC, **Ana Salgado**, Centro de Linguística da Universidade Nova de Lisboa (CLUNL) & Academia das Ciências de Lisboa, **Margarida Ramos**, CLUNL, **Sara Carvalho**, CLUNL & Centro de Línguas, Literaturas e Culturas, **Laurent Romary**, Automatic Language Modelling and ANAlysis & Computational Humanities Inria de Paris (ALMAnaCH), **Bruno Almeida**, CLUNL, **Mohamed Khemakhem**, ArcaScience, **Raquel Silva**, CLUNL, **Toma Tasovac**, Belgrade Center for Digital Humanities (BCDH), **Rute Costa**, CLUNL.

MorDigital

Introduction

- In this talk we will introduce the Portuguese national project **MORDigital** and present some updates on its current progress.
- MORDigital aims to make a historic Portuguese-language dictionary, *Diccionario da Lingua Portugueza* aka *Morais*, available as a digital resource.
- The project brings together some of the latest innovations in **computational lexicography**, the **digital humanities** and **linguistic linked data** including work on modelling lexicographic resources using standards such as **OntoLex**, **TEI Lex-0** and **LMF.**
- It also innovative in its creation and use of pipelines for converting retrodigitised dictionaries into computational lexical resources.

MorDigital

The *Diccionario da Lingua Portugueza*

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

MorDigital

The *Diccionario da Lingua Portugueza*

Frontispiece of *Morais* dictionaries (1789, 1813, 1823)


DICCIONARIO
DA
LINGUA PORTUGUEZA
COMPOSTO
PELO PADRE
D. RAFAEL BLUTEAU,
REFORMADO, E ACCRESCENTADO
POR
ANTONIO DE MORAES SILVA
NATURAL DO RIO DE JANEIRO.
TOMO PRIMEIRO.
A—K
LISBOA,
NA OFFICINA DE SIMÃO THADDEO FERREIRA.

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


DICCIONARIO
DA
LINGUA PORTUGUEZA
RECOPILADO
DOS VOCABULARIOS IMPRESSOS ATE' AGORA, E NESTA SEGUNDA EDIÇÃO NOVAMENTE EMENDADO, E MUITO ACCRESCENTADO,
POR
ANTONIO DE MORAES SILVA
NATURAL DO RIO DE JANEIRO.
OFFERECIDO
AO MUITO ALTO, E MUITO PODEROSO
PRINCIPE REGENTE N. SENHOR.
TOMO PRIMEIRO.
A—E.
LISBOA,
NA TYPOGRAPHIA LACERDINA.
ANNO DE 1813.

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


DICCIONARIO
DA
LINGUA PORTUGUEZA
RECOPILADO
DE TODOS OS IMPRESSOS ATE' O PRESENTE,
POR
ANTONIO DE MORAES E SILVA
NATURAL DO RIO DE JANEIRO.
OFFERECIDO
AO MUITO ALTO, E MUITO PODEROSO
SENHOR D. JOÃO VI,
TOMO PRIMEIRO.
A—F
LISBOA:
NA TYPOGRAPHIA DE M. P. DE LACERDA.
ANNO DE 1823.

Two volumes
A to K, 952 pp.
L to Z, 872 pp.

MorDigital

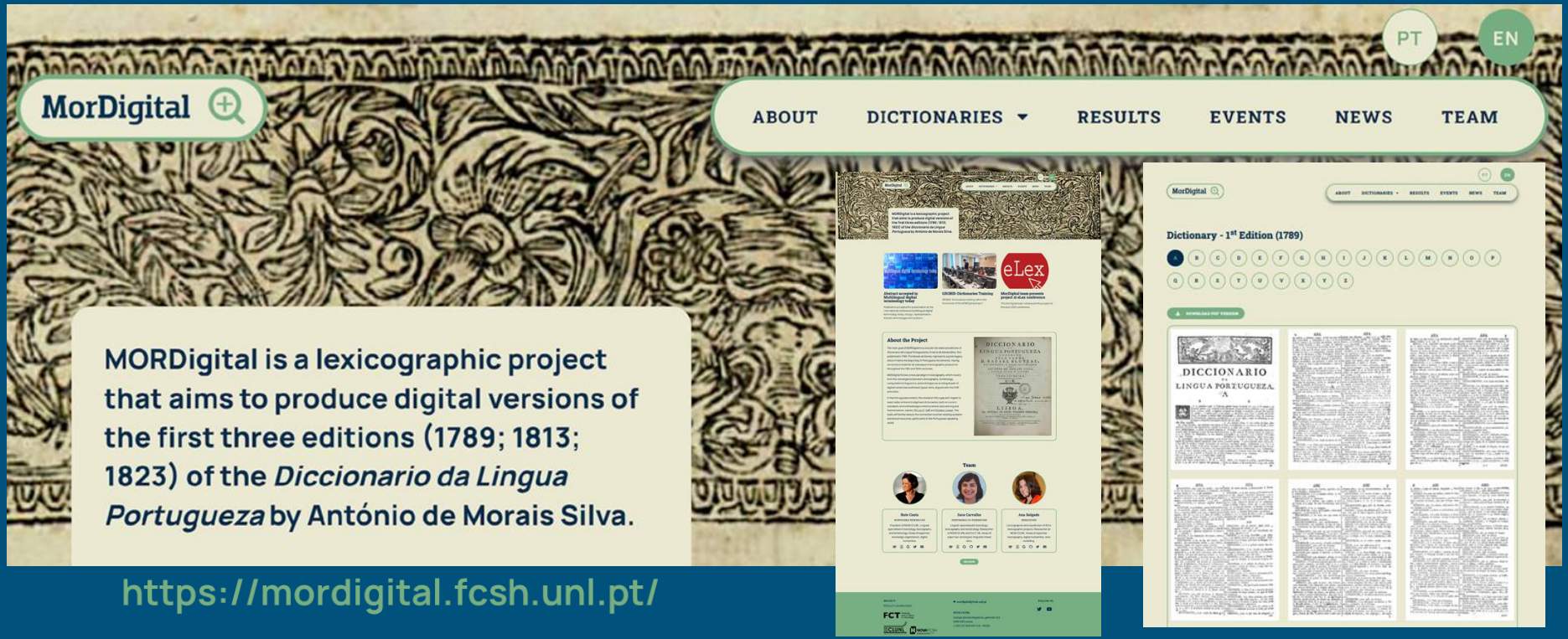
Who was Morais?

- The dictionary in question is usually referred to by the name of its compiler, the renowned Brazilian lexicographer **ANTÓNIO DE MORAIS SILVA** (1757?–1824)
- Born in Rio de Janeiro, Morais graduated in Civil and Canon Law from the University of Coimbra in Portugal.
- After being accused of heresy by the Inquisition, he fled to England and devoted himself to the study of languages; it was here that he planned the structure of his future dictionary
- Morais moved to Brazil in 1794 where he entered the judiciary and held the position of judge in the Bahia Court of Appeal (a position from which he soon resigned). He subsequently moved to Muribeca, in Pernambuco, where he lived until his death on April 11, 1824
- We can’t find any images of Morais himself!

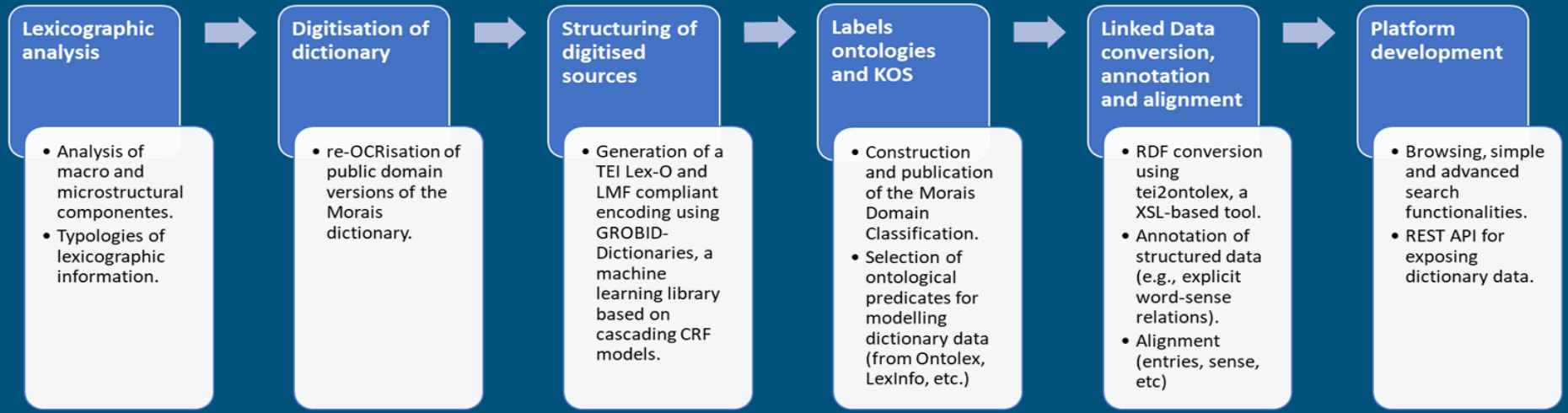
MORDigital

- **MORDigital – Digitalização do Diccionario da Lingua Portugueza de António de Morais Silva** [PTDC/LLT-LIN/6841/2020] is a project financed by the Portuguese National Funding agency through the FCT – Fundação para a Ciência e Tecnologia
- Although it is a Portuguese national project, it also includes collaborators from all over Europe.
- The main aim of the project is to convert **three editions** (1789; 1813; 1823) of *Morais* into a structured lexical resource in both TEI-XML (based on the ISO LMF standard) and in RDF (based on the OntoLex-Lemon model and its recent extensions).
- These editions will also be made available via an online interface on the website (at the moment only PDFs are available).

MORDigital - The website



Workflow


Lexicographic analysis
• Analysis of macro and microstructural componentes.
• Typologies of lexicographic information.
Digitisation of dictionary
• re-OCRisation of public domain versions of the Morais dictionary.
Structuring of digitised sources
• Generation of a TEI Lex-O and LMF compliant encoding using GROBID-Dictionaries, a machine learning library based on cascading CRF models.
Labels ontologies and KOS
• Construction and publication of the Morais Domain Classification.
• Selection of ontological predicates for modelling dictionary data (from Ontolex, LexInfo, etc.)
Linked Data conversion, annotation and alignment
• RDF conversion using tei2ontolex, a XSL-based tool.
• Annotation of structured data (e.g., explicit word-sense relations).
• Alignment (entries, sense, etc)
Platform development
• Browsing, simple and advanced search functionalities.
• REST API for exposing dictionary data.

Optical Character Recognition (OCR) Corrections

- We already had access to OCR'ed versions of the dictionary editions at the beginning of the project.
- These needed to be post-corrected. For this we decided to use **ABBYY FineReader**.
- An established tool with user-friendly interface.
- It creates editable, searchable files. Produces several output formats (RTF, DOCX, PDF, HTML, XML, etc.).
- The tool preserves typographical features.

Noise produced by the OCRization	Description	Example
Archaic characters	The **long /s/** is recognised as a lower-cased /f/ , or /j/ if the original is typed in italic	**f** instead of **ſ** **j** instead of **ʃ**
Case-sensitivity	Lower- and upper-case characters can be mixed up	**l** instead of **i**
Wrong characters inserted	Characters are misrecognised as wrong characters	**rn** instead of **m**
Segmentation errors	Different spacings between lines, words. Misrecognition of white-spaces	**temtres** instead of **tem tres**
Ligature	A combination of two or three characters set as a single glyph	**fi**
Ink spots	Text has ink spots, which prevents both human and machine from reading	

MorDigital

Structuring of Digital Editions

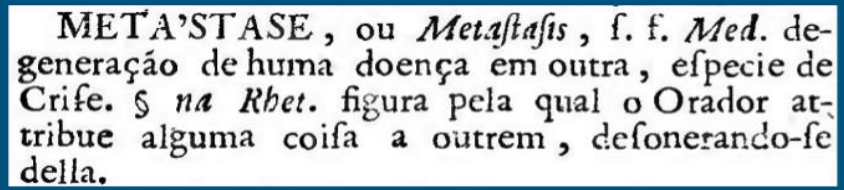
- For the structuring of the digital editions we are using **GROBID-Dictionaries**
- This is a machine learning library for structuring digitised lexical resources and entry-based documents with encyclopedic or bibliographic content.
- In particular, it allows the automatic parsing, extraction and structuring of lexical information from PDFs.
- GROBID-Dictionaries takes as input lexical resources digitised in PDF format and generates a TEI-encoded hierarchy of the different text structures which it has recognised.

MorDigital

TEI and OntoLex

- The TEI-XML sources will subsequently be converted to OntoLex (both the original model and its follow-up modules) using an XSLT stylesheet.
- Having an RDF version available allows us to make the dictionary available using both a dedicated platform and via a SPARQL endpoint.
- In addition, the RDF versions of each edition of the dictionary will be published on the LLOD cloud. This will be an important contribution to adding Portuguese language resources to the cloud.
- At the end of the project, based on our experiences, we will draft technical guidelines to help lexicographers and digital humanists carry out these tasks.

An Example Entry



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.

```
<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>
```

Organising knowledge

Morais	METALABEL	Encyclopedia of Diderot and d'Alembert	Dewey Decimal Classification (DDC)	EuroVoc	BabelNet
Arithm. Arithmetico	arithmetics	Reason Philosophy Science of Nature Mathematics Pure Arithmetics	500 Natural sciences & mathematics 510 Mathematics 513 Arithmetic	BT1 pure mathematics BT2 mathematics BT3 natural sciences	IS A 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	IS A 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	IS A Science major Universal language Academic discipline formal science

Costa et al. MDTT 2022

Organising knowledge

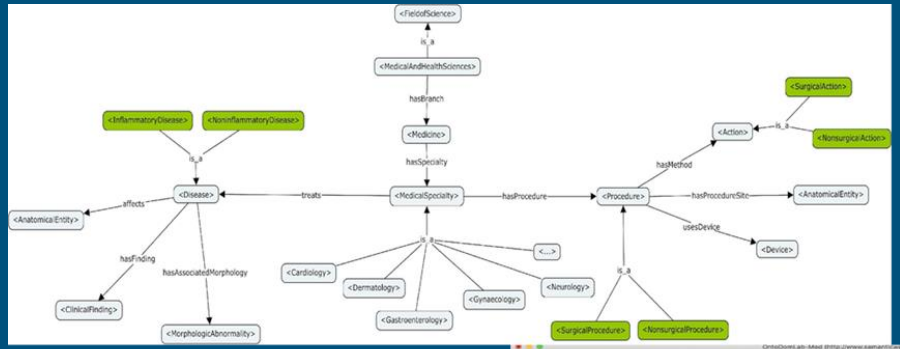
ARITHMETICA, ſ. f. arte de calcular por algariſmos.

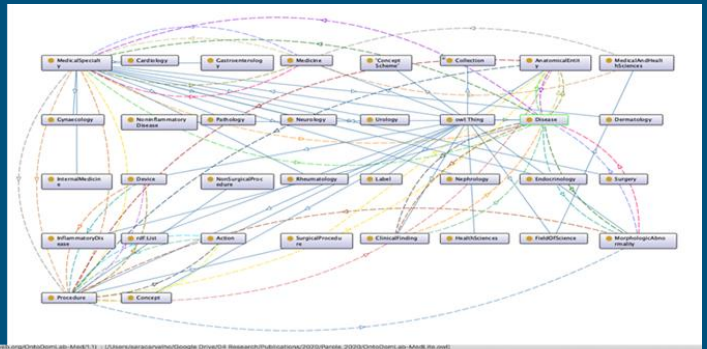
GEOMETRIA, ſ. f. parte da Mathematica que enſina a conhecer a grandeza, razões, e proporçóes das grandezas continuas, ou ſejão linhas, ou figuras, ou ſólidos, ou ſuperficies.

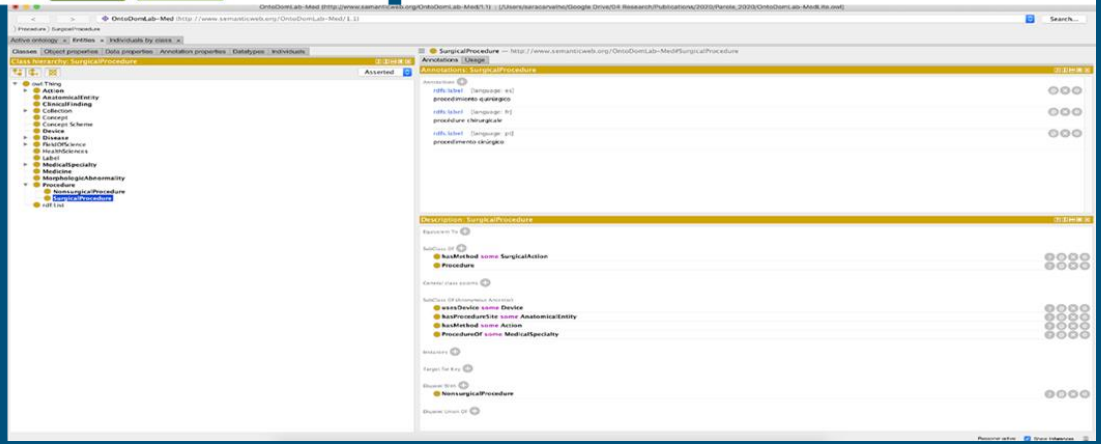
MATHEMATICA, ſ. f. a ſciencia, que enſina a conhecer as grandezas de toda ſorte, ſuas razões, relações, e proporçóes: *Mathematica miſta* (oppõe-ſe ás *puras*) a que enſina a applicar os principios de calculo, e geometria aos corpos.

MorDigital

Organising knowledge







OntoDomLab-Med: https://github.com/sarampcarvalho/OntoDomLab-Med

MorDigital

Ongoing Work

- Start OCR corrections on the two other editions of the dictionary
- Run the GROBID tool iteratively on the corrected output of the OCR of the first edition to ensure a correct TEI-XML (LMF) encoding of the different components of single entries (e.g., authoritative examples, collocations,)
- Start testing the XSLT transformation to OntoLex and its extensions
- In addition throughout the project our approach foresees a rigorous linguistic treatment of the source texts. This will make it possible to organise and structure the lexicographic components, and to elicit lexical relationships between various elements.
- We also propose combining semasiological and onomasiological approaches in our treatment of the different editions of Morais. For this we foresee the inclusion of ontologies (e.g. diasystematic marking, namely domain labels, registers and part of speech categories).

Thank You! Obrigado! Ačiū!

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

Ana Salgado: anasalgado@fcsh.unl.pt

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

Sara Carvalho: sara.carvalho@ua.pt

Laurent Romary: laurent.romary@inria.fr

Bruno Almeida: brunoalmeida@fcsh.unl.pt

Margarida Ramos: mvramos@fcsh.unl.pt

Mohamed Khemakhem: medkhemakhemfsegs@gmail.com

Toma Tasovac: ttasovac@humanistika.org

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

Toma Tasovac: ttasovac@humanistika.org