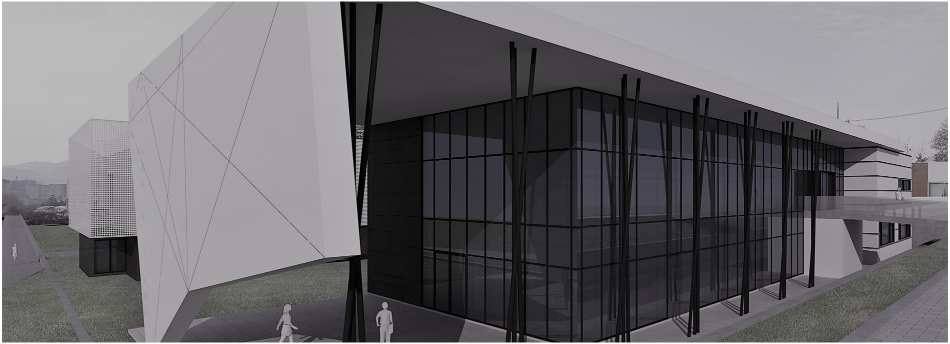
LeXmart

A Smart Tool for Lexicographers

A. Simões, A. Salgado, R. Costa & J. J. Almeida



IPCA
INSTITUTO POLITÉCNICO
DO CÁVADO E DO AVE

2Ai APPLIED
ARTIFICIAL
INTELLIGENCE
LABORATORY

CLUNL

Universidade do Minho
Escola de Engenharia

Outline

Motivation

LeXmart Birth

LeXmart

- Architecture
- End-User Tools
- Lexicographic Support Tools
- Content Management System
- RESTful API

Conclusions

Motivation

- Academy of Sciences Portuguese dictionary
- Published in paper format (2001)
- A new goal for 2019–202?: publish an updated version online
- The original dictionary was only available in PDF format
- A conversion to a computer readable format was required
- Need for a tool to allow concurrent edition of the dictionary

LeXmart Birth

- Develop a process to convert the PDF format into XML (TEI)
- Split the dictionary by entries, in different files
- Import entries into a database that:
  - should be aware of XML structure
  - should be able to query entries using XML structure
  - allow quick editing on XML aware IDEs
- eXist-DB was chosen
  (easy to integrate with Oxygen XML Editor)
- A set of small tools to assess current dictionary status:
  - validate entries
  - edit entries
  - show reports on entry classifications/grammar information

LeXmart

smart lexicography

http://lexmart.eu

Architecture

- A collection over eXist-DB for the dictionary:
  - Each dictionary entry is an individual XML file
  - Entries are codified using the TEI Dictionary module
  - TEI headers are not being currently stored

Architecture

- A collection over eXist-DB for the dictionary:
  - Each dictionary entry is an individual XML file
  - Entries are codified using the TEI Dictionary module
  - TEI headers are not being currently stored
- A collection for the application:
  - Development in XQuery 3.0 and XPath 3.0
  - TEI embedded into HTML directly, using CSS

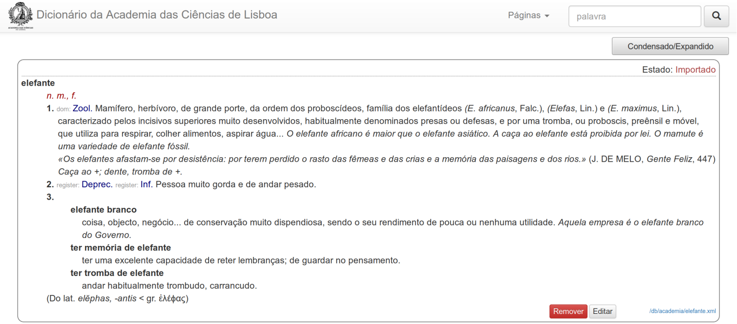
Architecture

- A collection over eXist-DB for the dictionary:
  - Each dictionary entry is an individual XML file
  - Entries are codified using the TEI Dictionary module
  - TEI headers are not being currently stored
- A collection for the application:
  - Development in XQuery 3.0 and XPath 3.0
  - TEI embedded into HTML directly, using CSS
- A collection for the CMS pages

End-User Tools

Headword search

- The typical use of a dictionary
- Quite efficient, querying the `orth` TEI elements.



Estado: Importado

elefante

n. m., f.

1. dom: Zool. Mamífero, herbívoro, de grande porte, da ordem dos proboscídeos, família dos elefantídeos *(E. africanus*, Falc.), *(Elefas*, Lin.) e *(E. maximus*, Lin.), caracterizado pelos incisivos superiores muito desenvolvidos, habitualmente denominados presas ou defesas, e por uma tromba, ou proboscis, preênsil e móvel, que utiliza para respirar, colher alimentos, aspirar água... *O elefante africano é maior que o elefante asiático. A caça ao elefante está proibida por lei. O mamute é uma variedade de elefante fóssil.*

«Os elefantes afastam-se por desistência: por terem perdido o rasto das fêmeas e das crias e a memória das paisagens e dos rios.» (J. DE MELO, *Gente Feliz*, 447)

Caça ao +; dente, tromba de +.

2. register: Deprec. register: Inf. Pessoa muito gorda e de andar pesado.

3.

elefante branco

coisa, objecto, negócio... de conservação muito dispendiosa, sendo o seu rendimento de pouca ou nenhuma utilidade. *Aquela empresa é o elefante branco do Governo.*

ter memória de elefante

ter uma excelente capacidade de reter lembranças; de guardar no pensamento.

ter tromba de elefante

andar habitualmente trombudo, carrancudo.

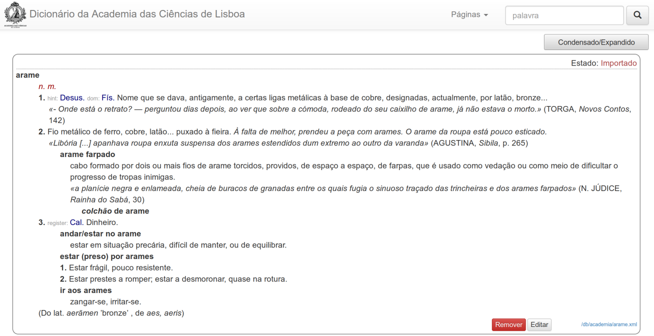
(Do lat. *elĕphas, -antis* < gr. ἐλέφας)

search by "elefante" (*elephant*)

End-User Tools

Reverse search

- not possible on paper dictionaries
- allow querying by the definition
- queries all entry definition information (slower)



Estado: Importado

arame

n. m.

1. hint: Desus. dom: Fís. Nome que se dava, antigamente, a certas ligas metálicas à base de cobre, designadas, actualmente, por latão, bronze...
«- Onde está o retrato? — perguntou dias depois, ao ver que sobre a cómoda, rodeado do seu caixilho de arame, já não estava o morto.» (TORGA, *Novos Contos*, 142)

2. Fio metálico de ferro, cobre, latão... puxado à fieira. *À falta de melhor, prendeu a peça com arames. O arame da roupa está pouco esticado.*
«Libória [...] apanhava roupa enxuta suspensa dos arames estendidos dum extremo ao outro da varanda» (AGUSTINA, *Sibila*, p. 265)

arame farpado

cabo formado por dois ou mais fios de arame torcidos, providos, de espaço a espaço, de farpas, que é usado como vedação ou como meio de dificultar o progresso de tropas inimigas.
«a planície negra e enlameada, cheia de buracos de granadas entre os quais fugia o sinuoso traçado das trincheiras e dos arames farpados» (N. JÚDICE, *Rainha do Sabá*, 30)

***colchão* de arame**

3. register: Cal. Dinheiro.

andar/estar no arame

estar em situação precária, difícil de manter, ou de equilibrar.

estar (preso) por arames

1. Estar frágil, pouco resistente.
2. Estar prestes a romper; estar a desmoronar, quase na rotura.

ir aos arames

zangar-se, irritar-se.

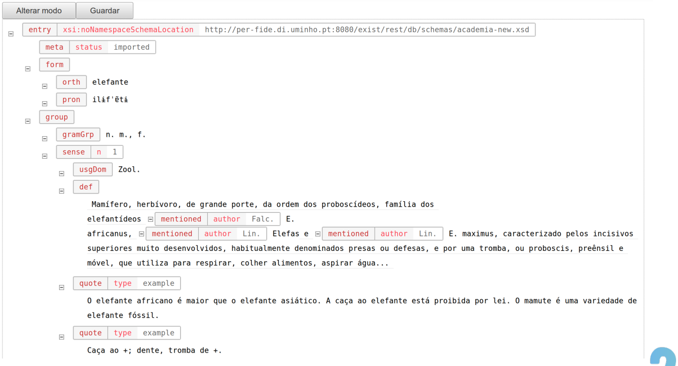
(Do lat. *aerāmen* 'bronze' , de *aes, aeris*)

search by “fio ferro” (*iron thread* → *wire*)

Lexicographic Support Tools

Entry editor

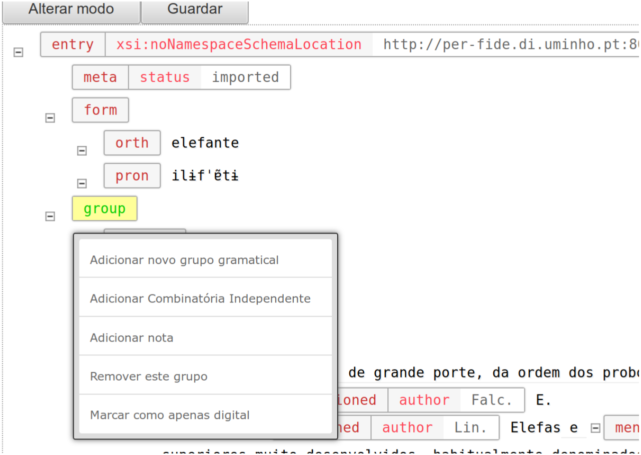
- Based on Xonomy XML editor (JavaScript based)
- Entries as fetched and saved directly from/into eXist-DB
- Configurable accordingly with the XML standard being used
- Contextual menus



```
⊟ entry  xsi:noNamespaceSchemaLocation  http://per-fide.di.uminho.pt:8080/exist/rest/db/schemas/academia-new.xsd
     meta  status  imported
  ⊟  form
    ⊟  orth  elefante
    ⊟  pron  ilɨf'ẽtɨ
  ⊟  group
    ⊟  gramGrp  n. m., f.
    ⊟  sense  n  1
      ⊟  usgDom  Zool.
      ⊟  def
            Mamífero, herbívoro, de grande porte, da ordem dos proboscídeos, família dos
           elefantídeos ⊟ mentioned  author  Falc.  E.
           africanus, ⊟ mentioned  author  Lin.  Elefas e ⊟ mentioned  author  Lin.  E. maximus, caracterizado pelos incisivos
           superiores muito desenvolvidos, habitualmente denominados presas ou defesas, e por uma tromba, ou proboscis, preênsil e
           móvel, que utiliza para respirar, colher alimentos, aspirar água...
      ⊟  quote  type  example
           O elefante africano é maior que o elefante asiático. A caça ao elefante está proibida por lei. O mamute é uma variedade de
           elefante fóssil.
      ⊟  quote  type  example
           Caça ao +; dente, tromba de +.
```

Lexicographic Support Tools

Entry editor


Alterar modo
Guardar
Adicionar novo grupo gramatical
Adicionar Combinatória Independente
Adicionar nota
Remover este grupo
Marcar como apenas digital

entry xsi:noNamespaceSchemaLocation http://per-fide.di.uminho.pt:8

meta status imported

form

orth elefante

pron ilɨf'ẽtɨ

group

de grande porte, da ordem dos probo

ioned author Falc. E.

ned author Lin. Elefas e men

superiores muito desenvolvidos, habitualmente denominado

Lexicographic Support Tools

Entry creation and deletion

- When viewing an entry, lexicographer can delete it (after confirmation)
- Adding a new entry will require the headword and:
  - guarantee no other entry with the same headword
  - create a stub XML document, adding it to the database
  - open the editor with that document

Lexicographic Support Tools

Meta Annotation

- Allow lexicographers to annotate entries and/or senses:
  - imported → directly imported from the paper dictionary
  - new → a new entry, created directly using LeXmart
  - edited → the entry was edited, but not yet considered final
  - revised → the entry can be made available publicly)

  (together with this information, a timestamp is recorded)
- Annotate entries as to be used only on digital version

Lexicographic Support Tools

Filters and Statistics

- Entries have diverse type of metainformation:
  - grammatical information
  - geographic source
  - register type
  - domain knowledge

Lexicographic Support Tools

Filters and Statistics

- Entries have diverse type of metainformation:
  - grammatical information
  - geographic source
  - register type
  - domain knowledge
- Filtering allows reports for specific kinds:
  - a list of terms from a specific geographic location
  - terms from a specific knowledge area
  - etc.

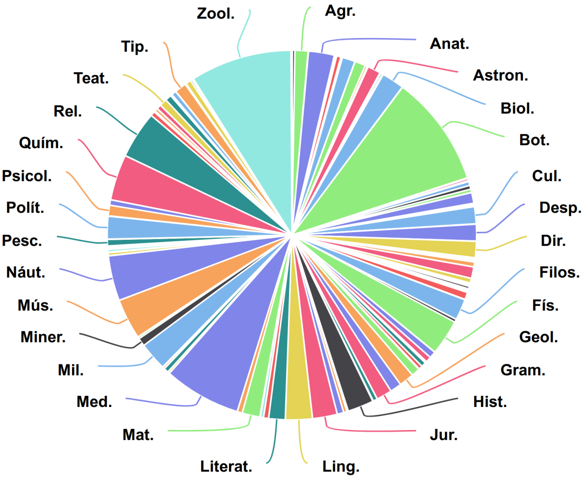
Lexicographic Support Tools

Filters and Statistics

- Entries have diverse type of metainformation:
  - grammatical information
  - geographic source
  - register type
  - domain knowledge
- Filtering allows reports for specific kinds:
  - a list of terms from a specific geographic location
  - terms from a specific knowledge area
  - etc.
- Statistics:
  - allow to evaluate relevance of classes
    How many entries in the cutlery domain?
  - Allow to validate possible duplicated tags

Lexicographic Support Tools

Filters and Statistics



Lexicographic Support Tools

Reports

- Allow to create custom reports on work evolution:
  - list of entries accordingly with their state (imported, new, edited, and revised)
  - digital only entries
  - entries from a specific defined list (given a text file with the list of terms)

Lexicographic Support Tools

Validation and Backup

- Validation:
  - LeXmart validates for well-formedness of the XML
  - LeXmart does not validate for DTD/Schema compliance in run time
  - Reports can be requested on Schema compliance

Lexicographic Support Tools

Validation and Backup

- Validation:
  - LeXmart validates for well-formedness of the XML
  - LeXmart does not validate for DTD/Schema compliance in run time
  - Reports can be requested on Schema compliance
- Backups:
  - eXist-DB backups are dumps of XML files
  - These files can be stored in a control version software
  - Using GIT allows us to have backups and version control easily

Content Management System

- Dictionaries are not just a list of definitions:
  - introduction
  - explanation of micro and macro structure
  - other decisions...
- Information to be shared by lexicographers:
  - rules when editing entries
  - general guidelines

Content Management System

- A simple CMS system was developed:
  - stored as independent collection on eXistDB
  - edited on a WYSIWYG editor (TinyMCE)
  - stored as XHTML
  - easy to display directly on a browser

RESTful API

- eXist-DB provides a RESTful API to:
  - fetch/store entries directly
  - query collections using XQuery
- On top of this, a generic user-centric API will be developed:
  - simple access to search/reverse search
  - filtering accordingly with entries metadata

Conclusions

- Solution based on Web-technologies
- Based on XML-aware database
- Extensible
- Available freely on GitLab (see `http://lexmart.eu`)
- Being used on a real dictionary project

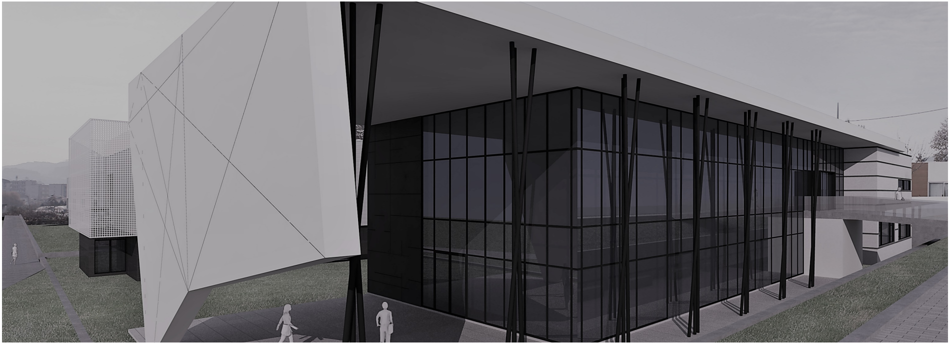
Future Work

- Translate the software
- Add configuration properties
- Stick to TEI and/or TEI Lex-0 standards
- Create “print-friendly” formats (XSL-FO or LaTeX)
- Integrate with other technologies:
    - ontologies and triple stores
    - wordnets

LeXmart

A Smart Tool for Lexicographers

A. Simões, A. Salgado, R. Costa & J. J. Almeida



INSTITUTO POLITÉCNICO DO CÁVADO E DO AVE

APPLIED ARTIFICIAL INTELLIGENCE LABORATORY

CLUNL

Universidade do Minho
Escola de Engenharia