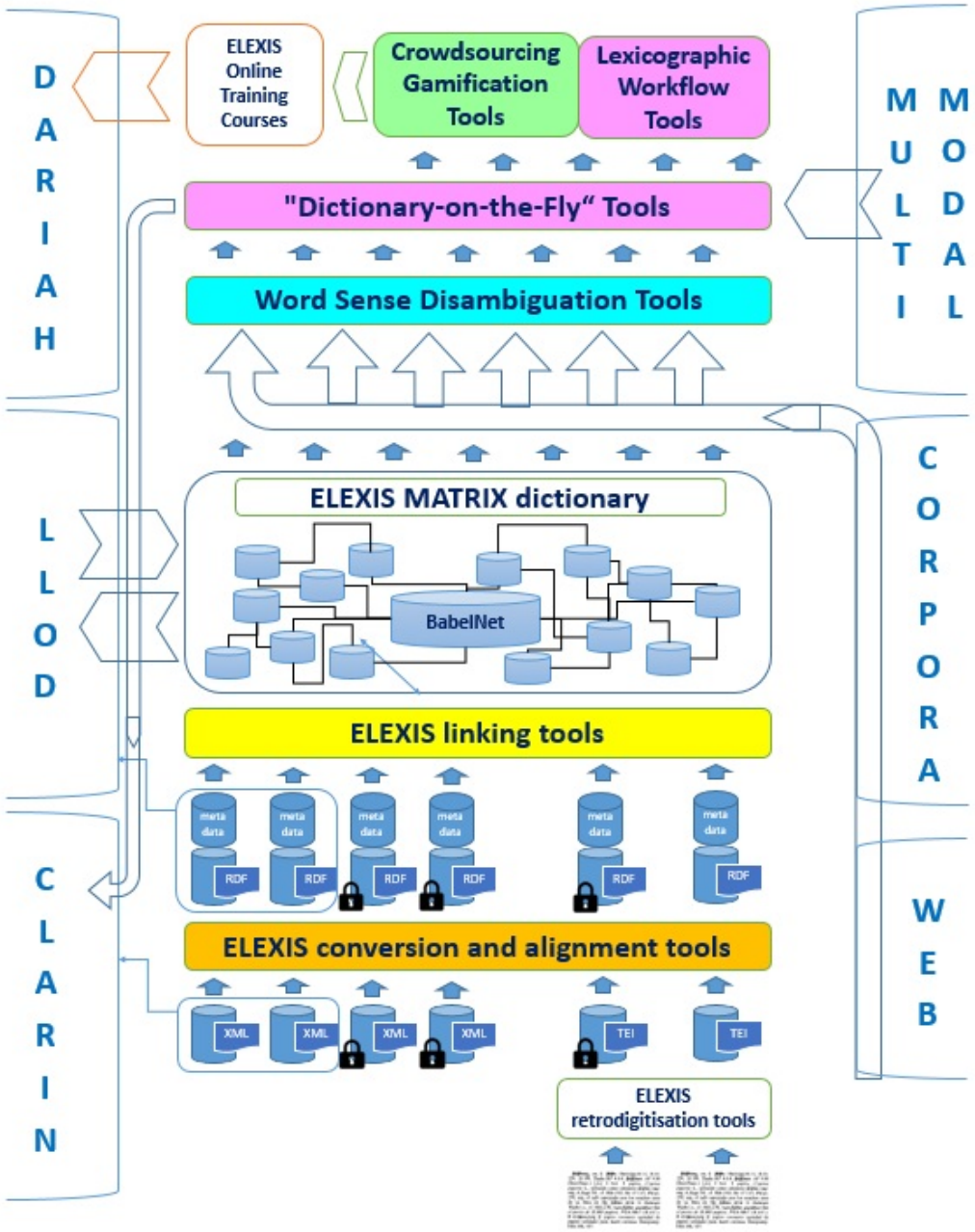
LEX 1 & LEX 3

Simon Krek (Jožef Stefan Institute, Ljubljana, Slovenia)

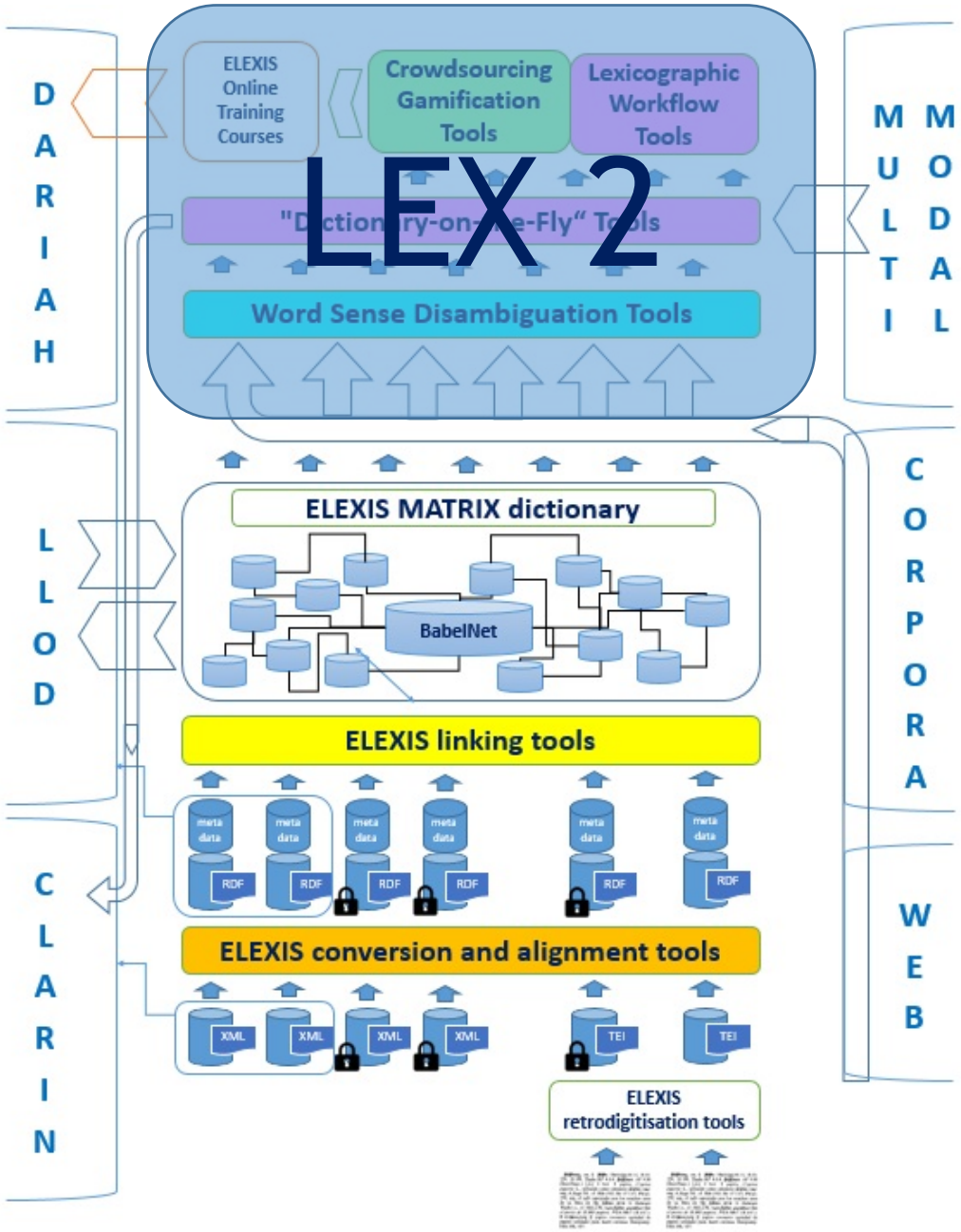
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure



VIRTUAL ACCESS

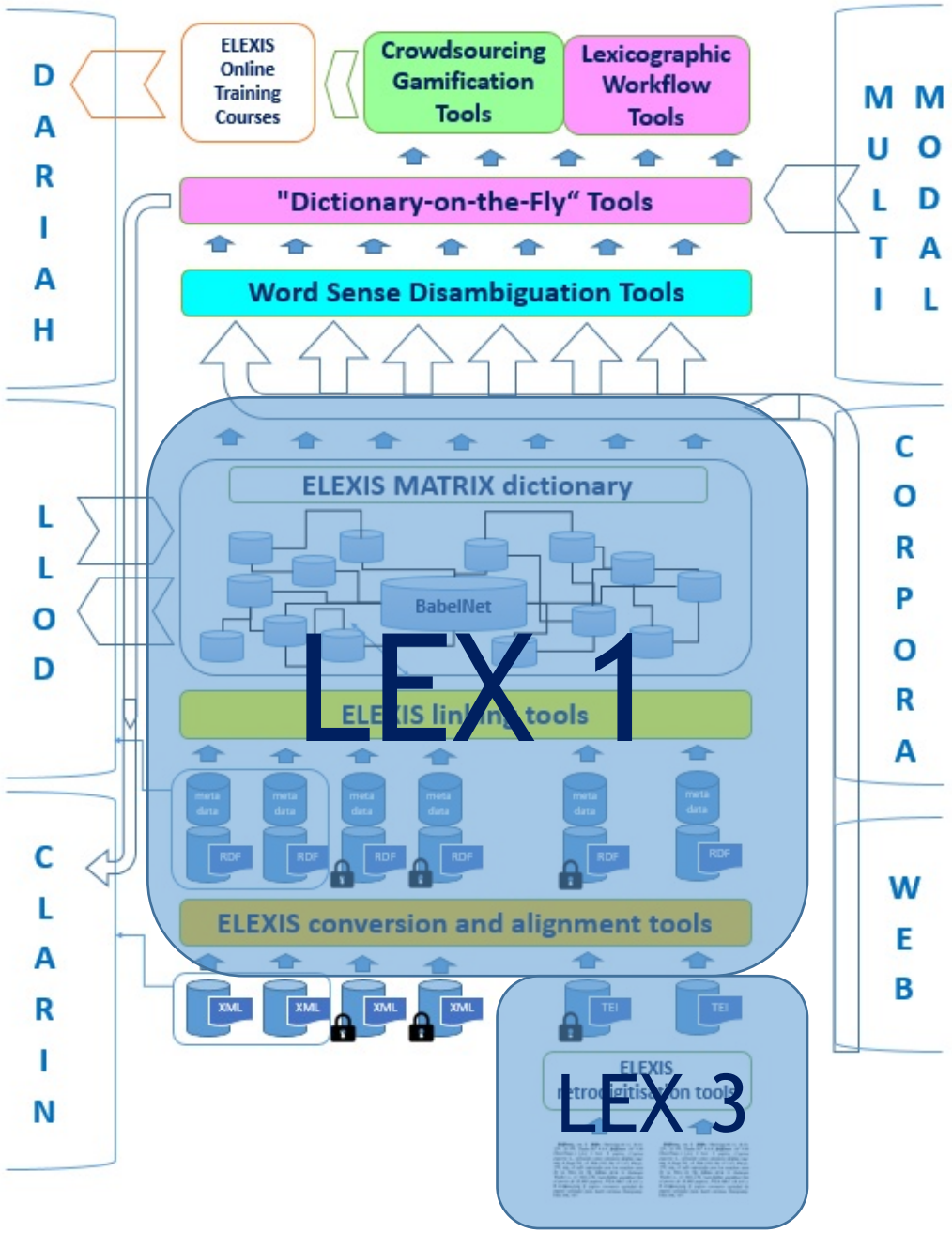
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.



elexis european lexicographic infrastructure

VIRTUAL ACCESS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.



elexis european lexicographic infrastructure

VIRTUAL ACCESS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

LEX 1 – segmentation, conversion tools

- The **LEX1 infrastructure** will be set up during the project and will consist of several sets of tools:
- (1) tools for automatic **segmentation and structuring** of lexicographic content for dictionaries that are currently produced in digital environments and are encoded in their own custom data format.
- (2) **conversion and alignment tools** that will provide users of the infrastructure with the possibility to harmonise and convert their lexicographic resources to a uniform data format that allows their seamless integration in Linked Open Data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

LEX 1 – linking tools

- (3) **linking tools** that will provide conceptual interoperability, enabling linking of ELEXIS lexicographic resources.
- This will provide the possibility to **link** lexical **entries**, **senses** and fundamental **concepts** in different lexical resources, using a semi-automatic approach.
- **BabelNet**, as an existing multilingual resource to provide cross-lingual linking, will be exploited for this purpose.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

LEX 1 – matrix dictionary = dictionary matrix

- (4) **ELEXIS matrix dictionary**: extensive linking of existing lexicographic resources by pivoting through BabelNet will enable the creation of what we call ELEXIS matrix dictionary –
- a universal **repository of linked senses**, meaning descriptions, etymological data, collocations, phraseology, translation equivalents, examples of usage and all other types of lexical information found in all types of existing lexicographic resources, monolingual, multilingual, modern, historical etc., available through a RESTful web service as **part of LEX1**.
- Freshly renamed: DICTIONARY MATRIX – **a matrix of dictionaries**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

LEX 3 infrastructure

- LEX3 infrastructure will be dedicated to retrodigitised dictionaries and will include
- (1) tools for automatic **segmentation and structuring** of content in retro-digitised dictionaries, and
- (2) **Legacy Dictionary Viewer**: a generic, modular dictionary publication tool for retrodigitised dictionaries which offers interfaces for the analysis and profiling of the underlying lexical data.
  - The viewer will allow users — lexicographers and digital humanists alike — to publish user-friendly, online editions or retrodigitised dictionaries without facing extensive technical or financial obstacles.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Segmentation and structuring

LEX 1 scenario: digital, unstructured

dictatorial /dɪktətɔːrɪəl/. ADJ Kdor je **dictatorial**, svojo oblast uporablja preveč nasilno. ♦ diktatorski, tiranski; *...dictatorial regimes. ...their dictatorial attitude.*
dictatorship /dɪktɛɪtə^{s}ʃɪp/, dictatorships. **1** UNCOUNT N **Dictatorship** je vladavina diktatorja. ♦ diktatura; *Democracy soon gave way to dictatorship.* **2** COUNT N **A dictatorship** je država, ki ji vlada diktator. ♦ diktatorska država
diction /dɪkʃə^{0}n/. UNCOUNT N **One's diction** je jasnost, s katero nekdo govori ali poje. ♦ način izražanja, izgovorjava, dikcija
dictionary /dɪkʃənə^{0}rɪ/, dictionaries. COUNT N **A dictionary** je knjiga, v kateri je seznam besed nekega jezika po abecednem redu, poleg njih pa so navedeni pomeni ali ustrezne besede v drugem jeziku. ♦ slovar; *...an English-French dictionary.*
did /dɪd/. **Did** je preteklik glagola TO DO. ♦ storiti
didactic /dɪ^{5}dæktɪk/. ADJ Kar je **didactic**, naj bi nas nekaj naučilo; knjižna raba. ♦ poučen

differ from the early ones in many ways. **2** VB Kadar ljudje **differ about something**, se o nečem ne strinjajo med seboj. ♦ ne strinjati se; *We differ about moral standards.*
difference /dɪfə^{0}rəns/, differences. **1** COUNT N or UNCOUNT N **The difference between things** je način, po katerem se stvari med seboj razlikujejo. ♦ razlika, razloček; *There is an essential difference between computers and humans. Look at their difference in size.* **2** SING N **The difference between two amounts** je količina, za katero je en znesek manjši ali večji od drugega. ♦ razlika **3** COUNT N Kadar ljudje **have their differences**, se z nečim ne strinjajo. ♦ nestrinjanje, spor **4** PHRASE Kadar nekaj **makes a difference**, spremeni neki položaj. ♦ biti pomembno; *It makes no difference whether he is a citizen or not.*
different /dɪfə^{0}rənt/. **1** ADJ Če je ena stvar **different from another**, je v določeni lastnosti drugačna od druge stvari. ♦ razlikovati se; *The meeting was different from any that had gone before. His message is very different to theirs.*

LEX 3 scenario: retro-digitised

ABERDEEN [*abᵉrdin'*], v. d'Ecosse, ch.-l. de comté; port sur la mer du Nord; 170.000 h. Université.
ABERDEEN (G. H. Gordon, *comte d'*), homme d'Etat anglais, né à Edimbourg. Premier ministre en 1852, il conclut avec la France une alliance contre la Russie (1784-1860).
ABER-VRACH, fl. côtier du Finistère (Atlantique); 34 kil. Station marémotrice d'essai.
ABGAR, nom de huit rois d'Edesse, en Mésopotamie (132 av. J.-C.-216 apr.).
ABIA, roi de Juda, fils de Roboam, vainqueur de Jéroboam, roi d'Israël (957-955 av. J.-C.).
ABIDJAN, ch.-l. de la Côte-d'Ivoire (A.-O. F.), sur une vaste lagune navigable; 15.000 h.
ABIMÉLECH [*lèk*], fils de Gédéon. Il devint Juge d'Israël, après avoir fait égorger ses frères; il établit son pouvoir sur Sichem et fut tué au siège de Thèbes, en Palestine (vers 1100 av. J.-C.).
ABIRON, lévite qui fut englouti dans la terre avec Coré et Dathan, tous trois révoltés contre Moïse et Aaron (*Bible*).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

GROBID-dictionaries

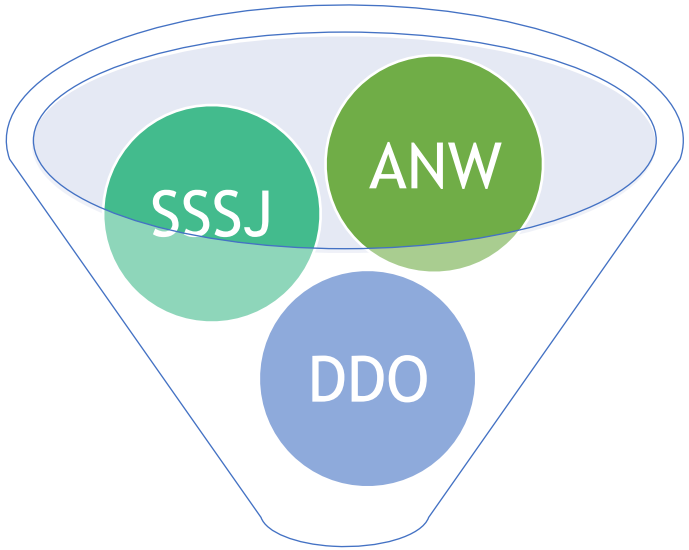
- GROBID-Dictionaries is a GROBID sub-module, implementing a java machine learning library, for **structuring digitized lexical resources** and entry-based documents with encyclopedic or bibliographic content.
- It allows the parsing, extraction and structuring of text information in such resources.

- Presentation: Mohamed Khemakhem
- On GitHub: https://github.com/MedKhem/grobid-dictionaries
- ELEXIS installation: http://grobid.elex.is/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

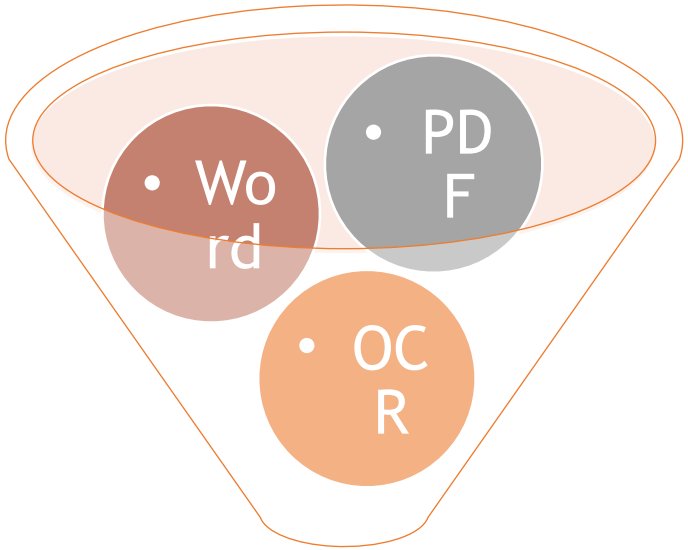
Conversion and alignment

ELEXIS (XML schema) conversion tools



TEI-Lex0

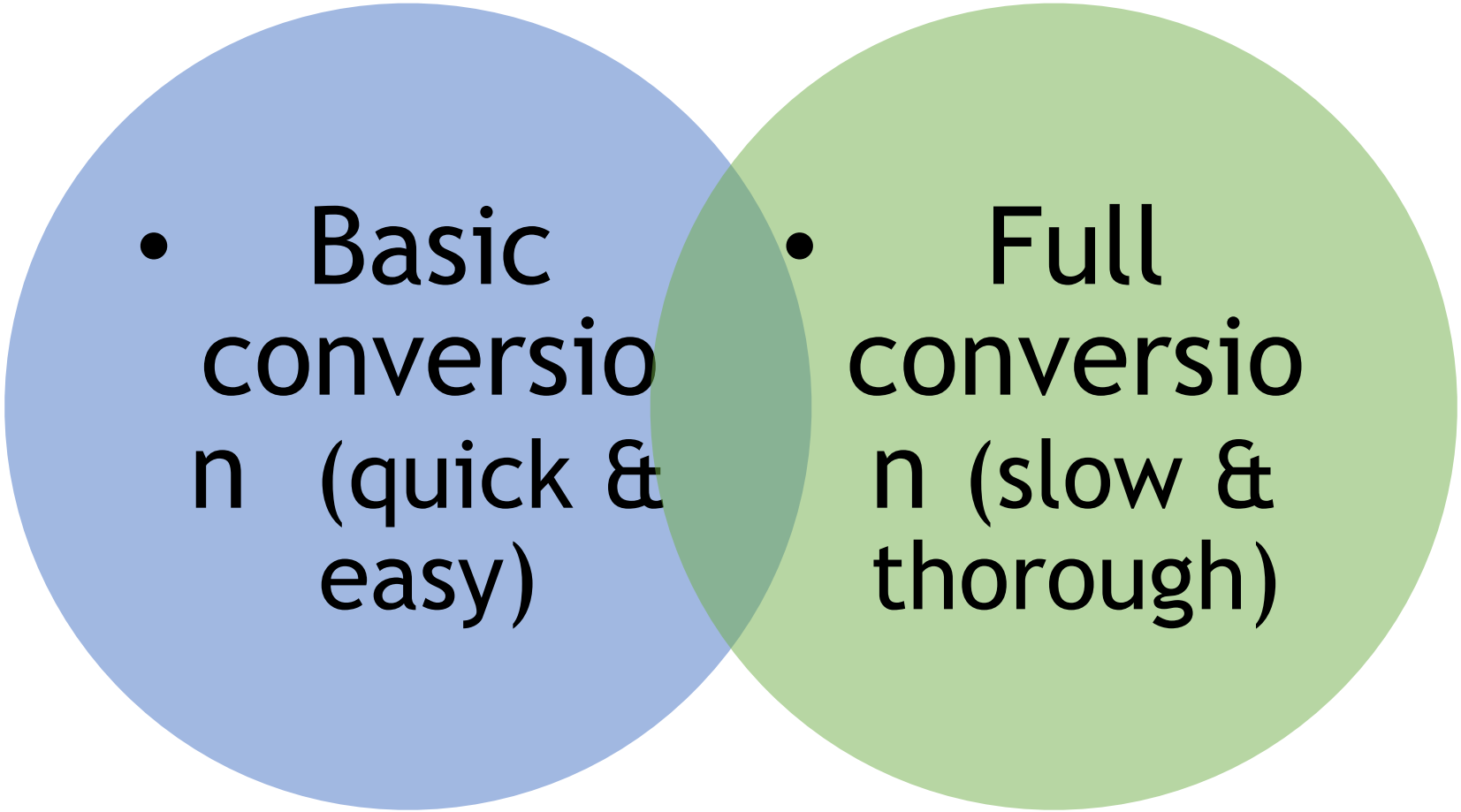
GROBID-dictionaries



- TEI-Lex0

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Conversion and alignment – TEI-Lex0



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Conversion tools - process

- Quick and easy
  - Upload your dictionary in custom XML format -> parse (or provide schema)
  - **Identify basic dictionary elements defined by ELEXIS data model**
  - Convert your dictionary to TEI-Lex0 and download
- Slow and thorough
  - Upload your dictionary in custom XML format -> parse (or provide schema)
  - **Align your schema and TEI-Lex0**
  - Convert your dictionary to TEI-Lex0 and download

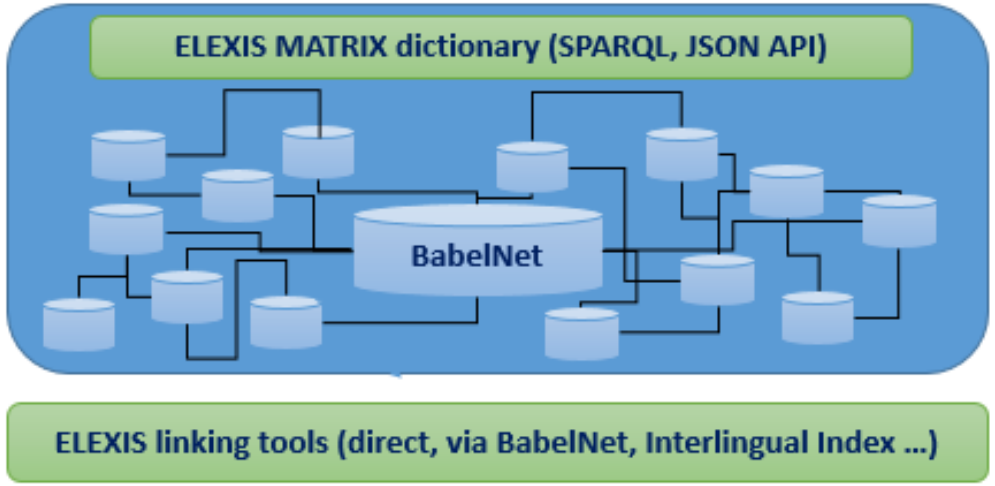
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

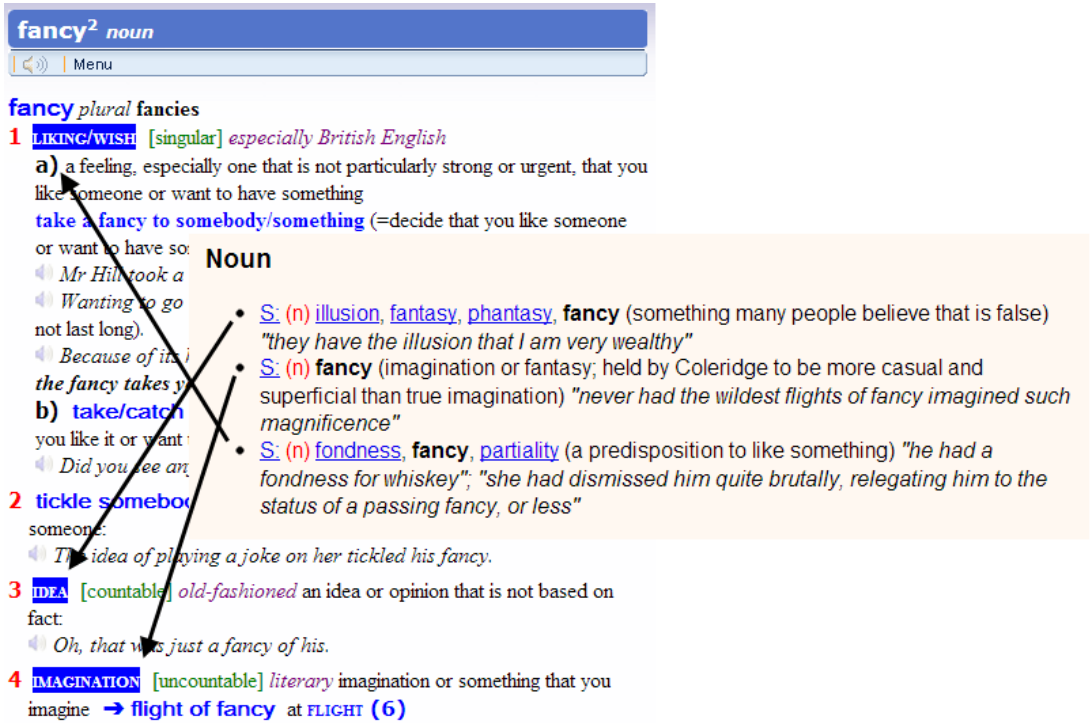
Conversion and alignment tools

- **(E)LEXIFIER**
  - *Lexifier - the dominant language of a pidgin or creole that serves as the basis for most of its vocabulary (Wiktionary)*
- The tool identifies only basic dictionary elements (entry, headword, sense, definition, example etc.) and produces a file in valid TEI-Lex0 format with original dictionary content.
- XSLT tools for “complete” conversion

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.

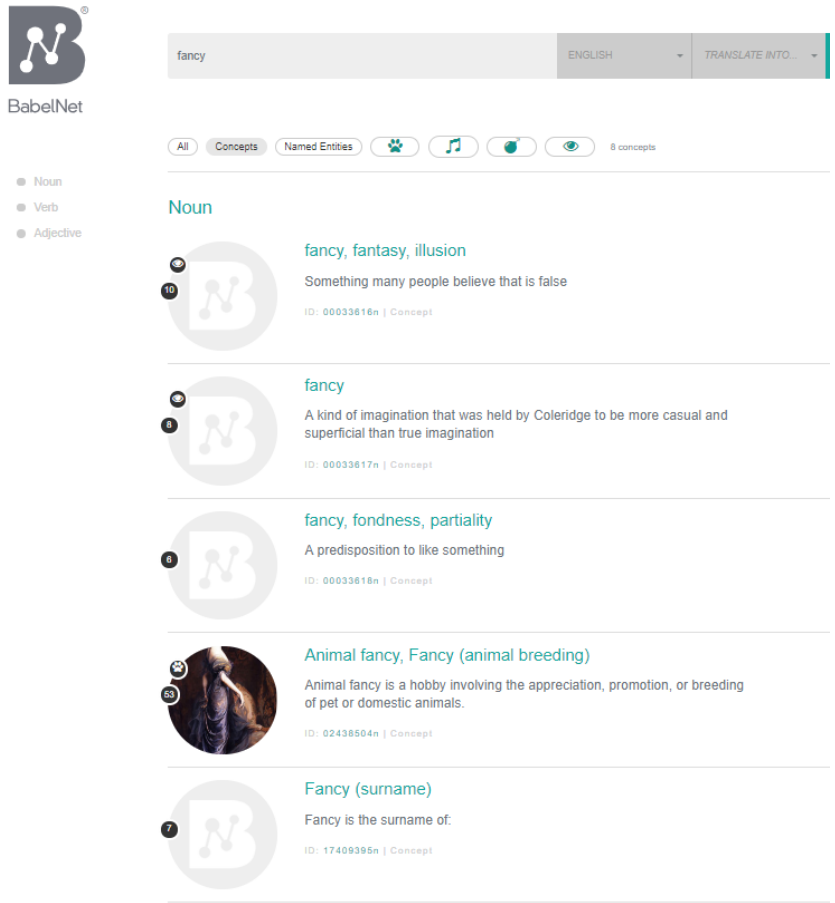
ELEXIS DICTIONARY MATRIX

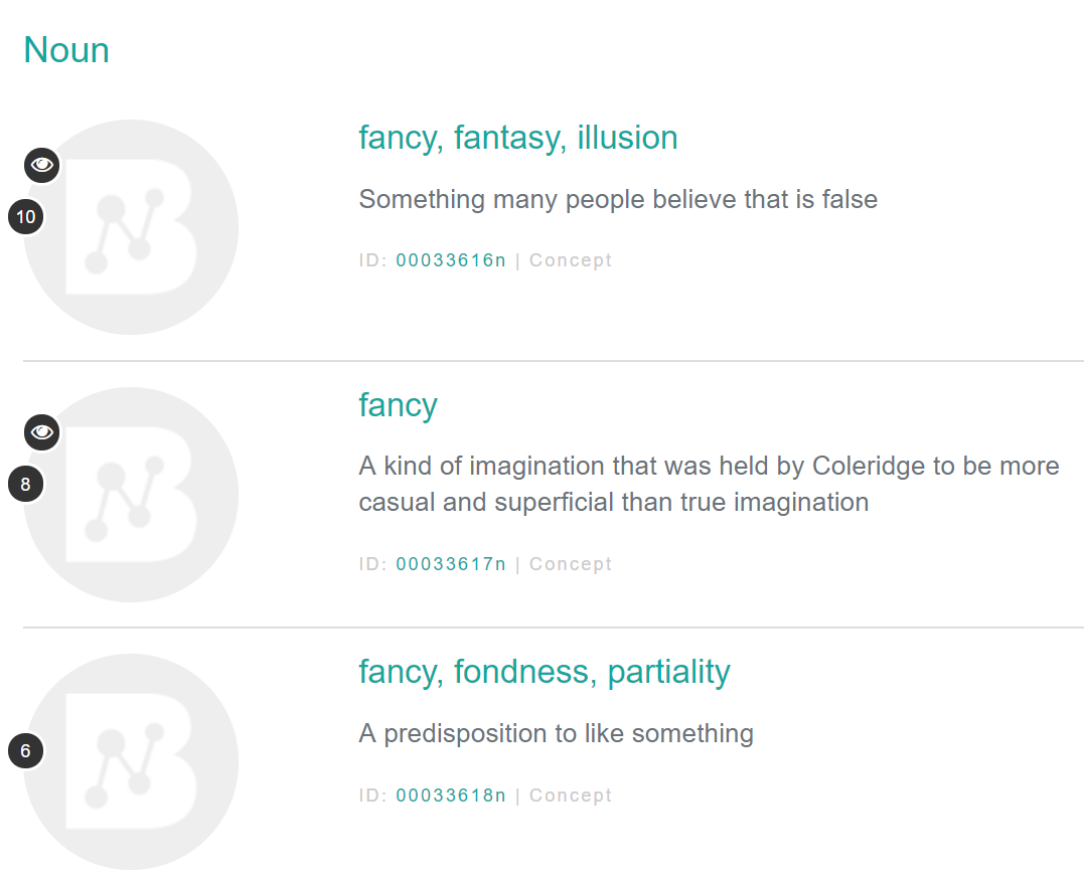




This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

BabelNet – five concepts FANCY-n

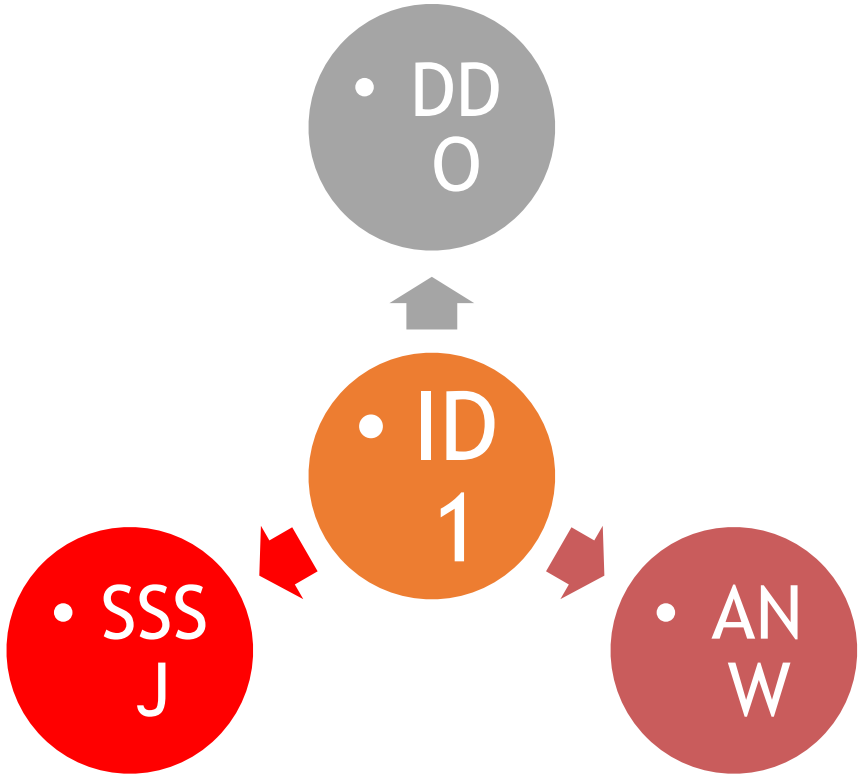




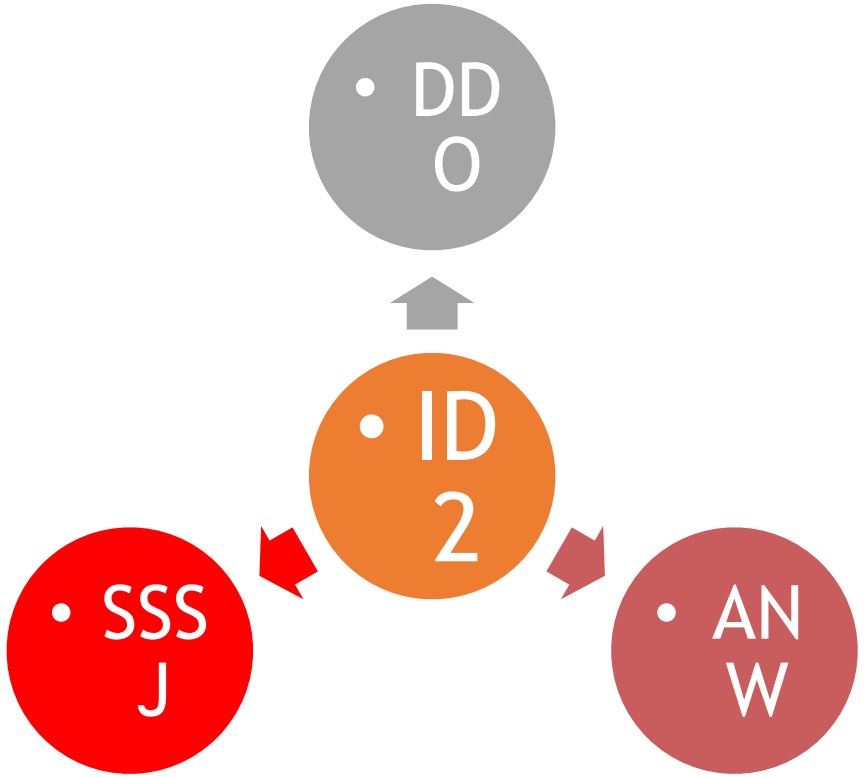
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Repository of senses – “matrix dictionary”

CONCEPT 1 - “FANCY-n”: something many people believe is false



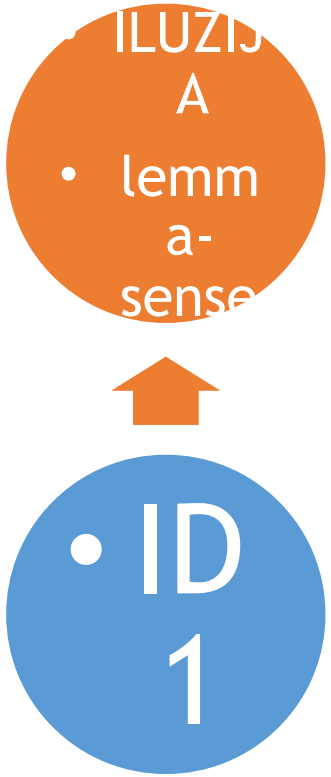
CONCEPT 2 “FANCY-n” CONCEPT 2: a predisposition to like something



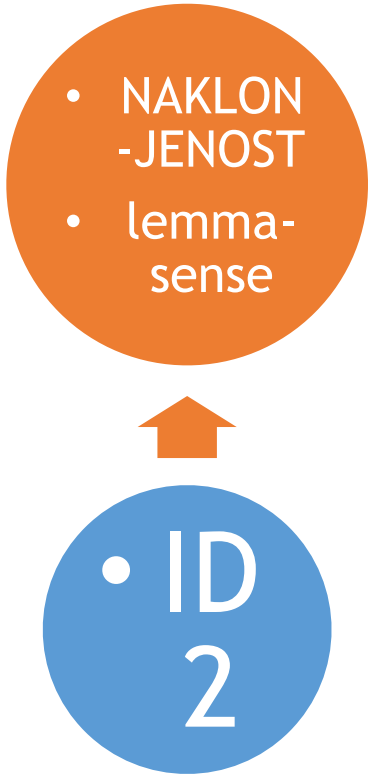
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.

Repository of senses - SloWNet

CONCEPT 1 - “FANCY-n”: something many people believe is false



CONCEPT 2 “FANCY-n” CONCEPT 2: a predisposition to like something



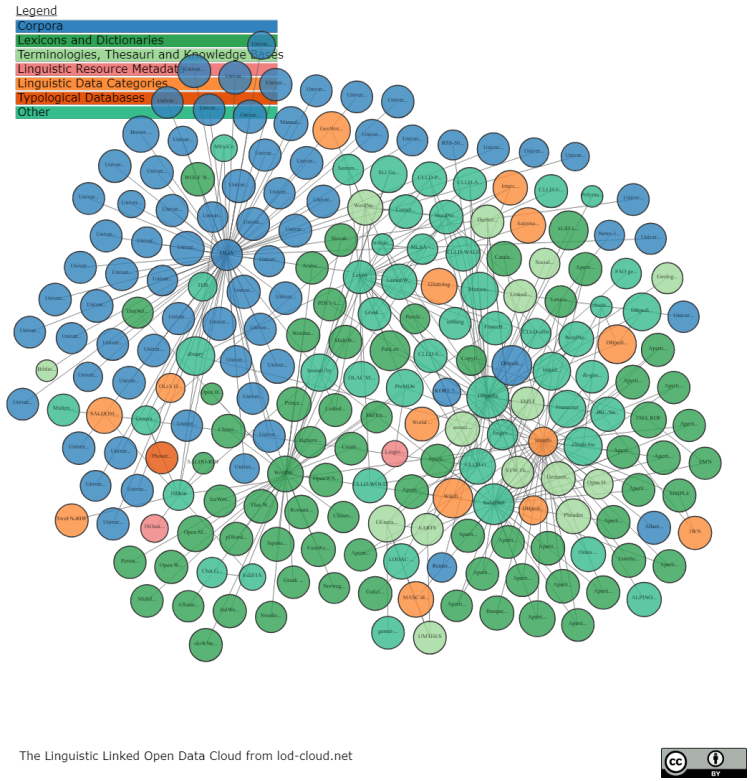
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure

EDM visibility

European Dictionary Portal (EDP)

Linguistic Linked Open Data cloud (LLOD)



The Linguistic Linked Open Data Cloud from lod-cloud.net

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Presentations

- GROBID-Dictionaries
- ELEXIS data model
- Linking & dictionary matrix
- Word Sense Disambiguation

- **Mohamed Khemakhem**
  - Inria-ALMAnaCH
- **Carole Tiberius**
  - Dutch Language Institute
- **John McCrae**
  - National University of Ireland
- **Roberto Navigli**
  - Sapienza University of Rome

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Timeline – LEX1 / 3

- **T1.4** Data Structuring for existing resources: **M1-M24**
  - GROBID-dictionaries, ELEXIFIER
    - **D1.3** Tools for the automatic segmentation and identification of lexicographic content (**M24**) (software)
    - **D1.4** ELEXIS conversion tools (**M36**) (software)
- **T2.2** Semi-automatic linking of lexical resources: **12M-48M**
  - just started → end of project
    - **D2.2** Interoperable interface for Lemon and TEI resources (**M24**) (software)
    - **D2.4** Cross-lingual Lexical Resource Linking Web Service (**M42**) (software)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.