ELEXIS project overview

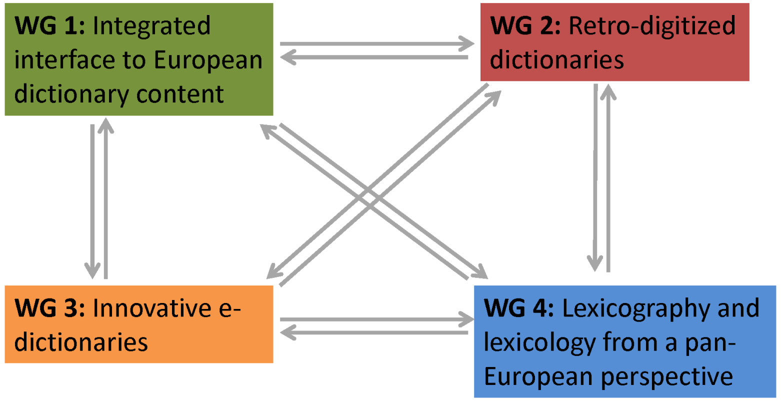
Simon Krek (Jožef Stefan Institute, Ljubljana, Slovenia)

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

European Network of e-Lexicography (ENeL)

Kick-off meeting, 11 October 2013, Brussels

Working Groups



WG 1: Integrated interface to European dictionary content

WG 2: Retro-digitized dictionaries

WG 3: Innovative e-dictionaries

WG 4: Lexicography and lexicology from a pan-European perspective

Observations

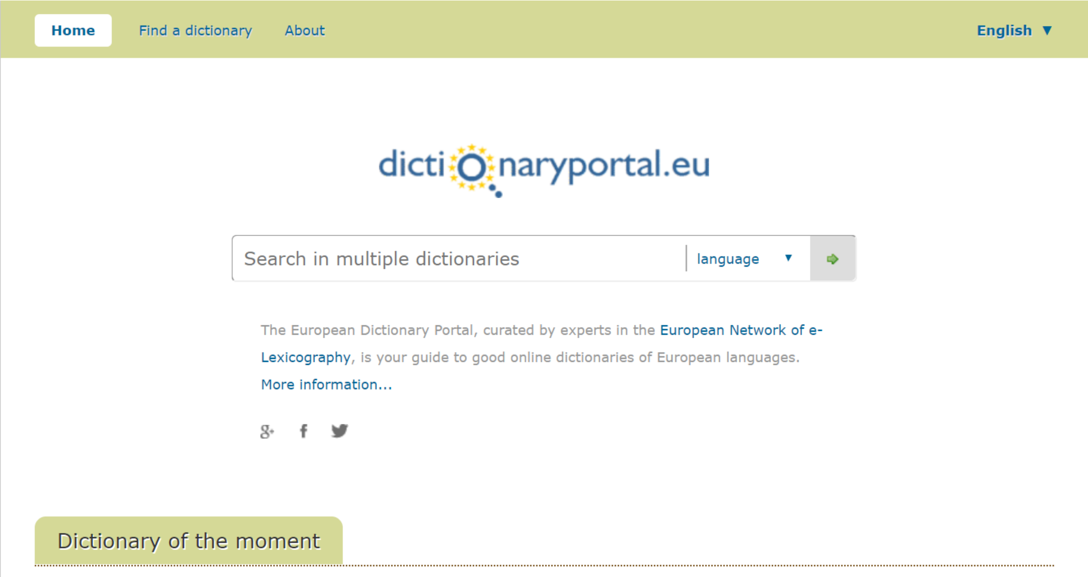
- Prevalence of user-generated dictionaries/growing gap between scholarly dictionaries and the general public;
- Lack of common standards and solutions for retrodigitized dictionaries;
- Lack of a common research paradigm, common standards and solutions for e-lexicography;
- In dictionaries languages are treated as isolated entities.

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

european lexicographic infrastructure

European Network of e-Lexicography (ENeL)

Final meeting, 18 September 2017, Leiden



The European Dictionary Portal, curated by experts in the European Network of e-Lexicography, is your guide to good online dictionaries of European languages.
More information...

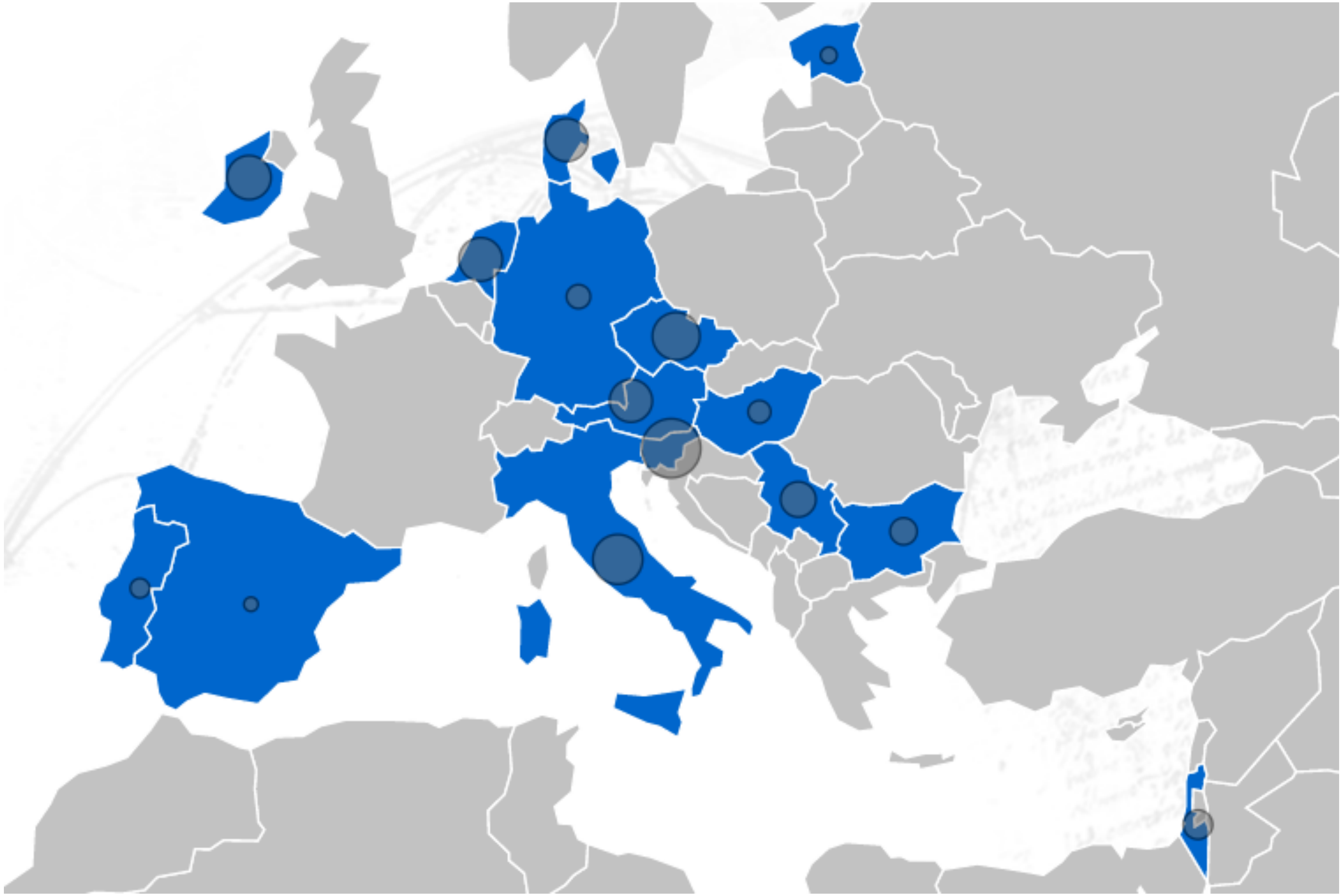
- 280 members
- 30 countries
- 7 meetings
- 4 workshops
- 37 STSMs
- 3 training schools
- 1 portal
- > 1 EU-funded project

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

ELEXIS FACT SHEET

- Call & Topic: INFRAIA-02-2017
  - Integrating Activities for **Starting Communities** (Publication: October 2015)
  - Model: two-stage (March 2016, March 2017), results: August 2017
- Start date: **1 February 2018**
- Duration: **48 months** (31 January 2022)
- Total cost: 4,999,967.50 €
- Coordinator: Jožef Stefan Institute, Ljubljana, Slovenia
- Number of partners: 17 from 15 countries
- Web site: www.elex.is

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



- Partners with:
- lexicographic data and/or expertise
- computational linguistics data and/or expertise
- expertise in standardisation
- digital humanities partners
- technology partners

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

GOALS

- To integrate, extend and harmonise national and regional efforts in the **field of lexicography**,
  - both modern and historical,
- with the goal of creating a sustainable **infrastructure** which will
  - (1) enable efficient **access to** high quality **lexical data** in the digital age, and
  - (2) **bridge the gap** between more advanced and lesser-resourced scholarly communities working on lexicographic resources.

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

EXPECTED IMPACTS

- Providing efficient **access to** quality lexicographic **data**
- Enabling massive **integration of** knowledge-based **resources**
    - Facilitating inclusion of innovative lexicography in **research** and **education**
    - Enabling the use of new technology and data in **industry**
    - Establishing **inter-infrastructure** synergies and optimisation
- RESULT: a **new** type of **lexicography** that no longer views languages as isolated entities

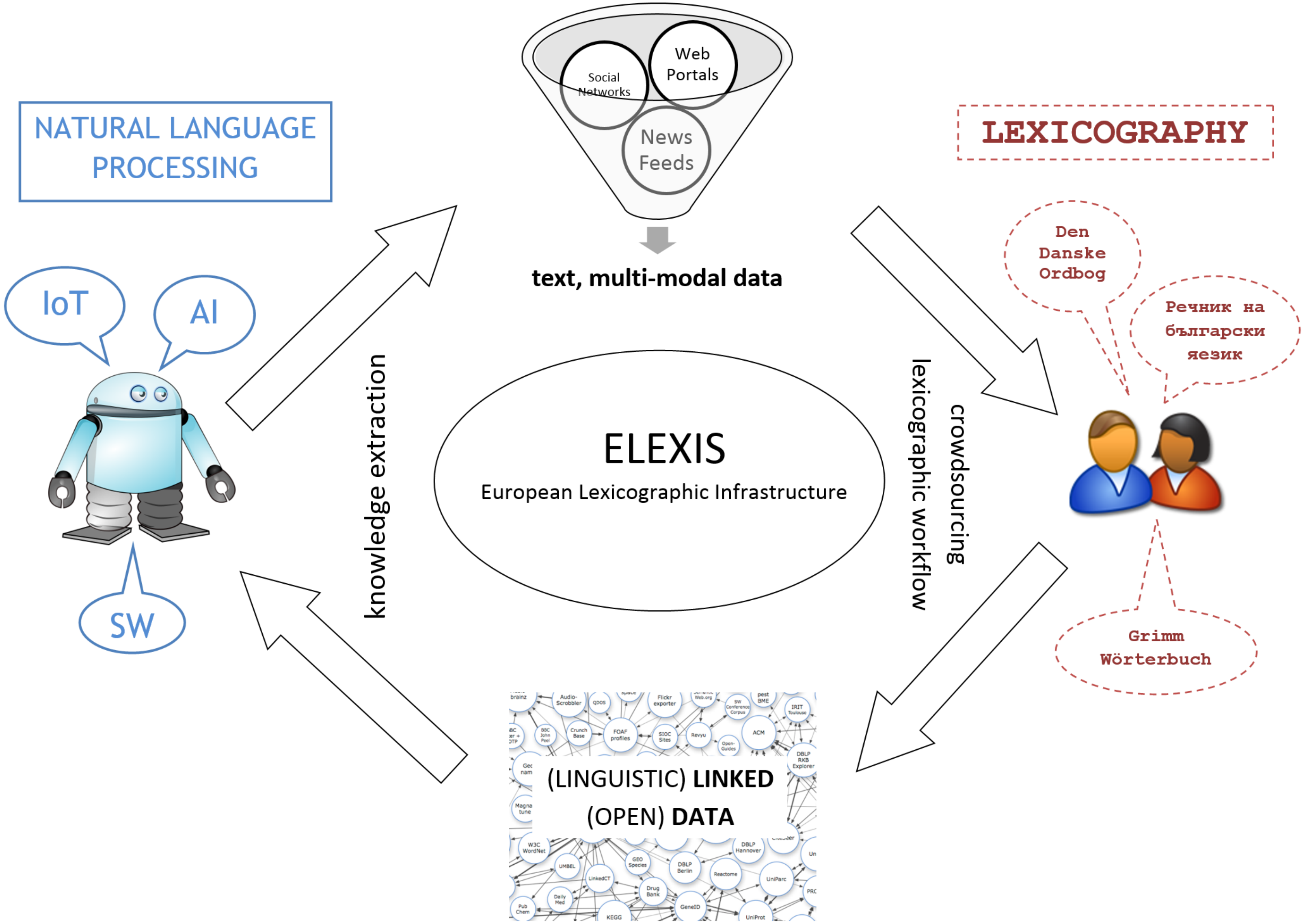
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

VIRTUOUS CIRCLE/CYCLE OF E-LEXICOGRAPHY

- “In the best of all possible worlds, computational enhancement and lexicographical upgrading would build upon each other in a virtuous circle that knew no bounds”.
    - Abstract: “NLP needs dictionaries, and dictionary-makers can use NLP to make better dictionaries, so there is great potential for synergy between the two activities. To date, there has been only very limited collaboration. This is substantially owing to dictionary publishers‘ concerns regarding intellectual property. In this paper I explore the different interests of publishers and NLP researchers, and present a business model which pays heed to both.”
  - Kilgarriff, A. (2000). Business models for dictionaries and NLP. International Journal of Lexicography, 13(2):107–118.

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



NATURAL LANGUAGE PROCESSING

Social Networks

Web Portals

News Feeds

text, multi-modal data

LEXICOGRAPHY

Den Danske Ordbog

Речник на българския език

Grimm Wörterbuch

IoT

AI

SW

knowledge extraction

ELEXIS

European Lexicographic Infrastructure

crowdsourcing

lexicographic workflow

(LINGUISTIC) **LINKED** (OPEN) **DATA**

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

THREE TYPES OF ACTIVITIES

- Joint **research** activities
  - tools and methods for enabling linked lexicographic resou
  - tools and methods to support innovative e-lexicography
- **Trans-national access** or **virtual access** activities
- **Networking** activities
  - documentation, guidelines, collections of best practices
  - online training modules
  - data seal of compliance for lexicographic data
  - workshops, seminars, and conferences
  - international forum

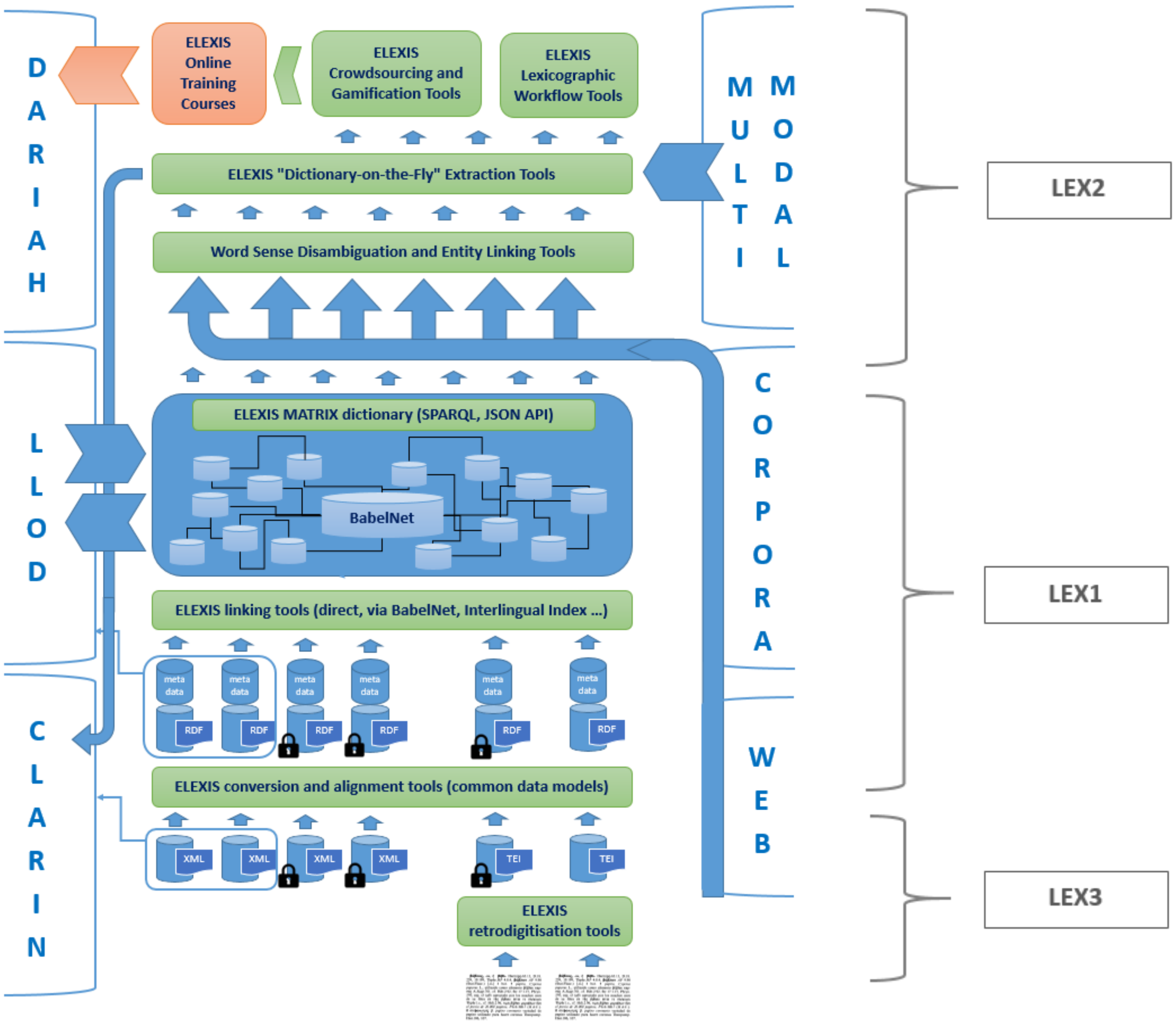
MONDAY
February 18

Technology day

TUESDAY
February 19

Community building day

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



ELEXIS Online Training Courses

ELEXIS Crowdsourcing and Gamification Tools

ELEXIS Lexicographic Workflow Tools

ELEXIS "Dictionary-on-the-Fly" Extraction Tools

Word Sense Disambiguation and Entity Linking Tools

ELEXIS MATRIX dictionary (SPARQL, JSON API)

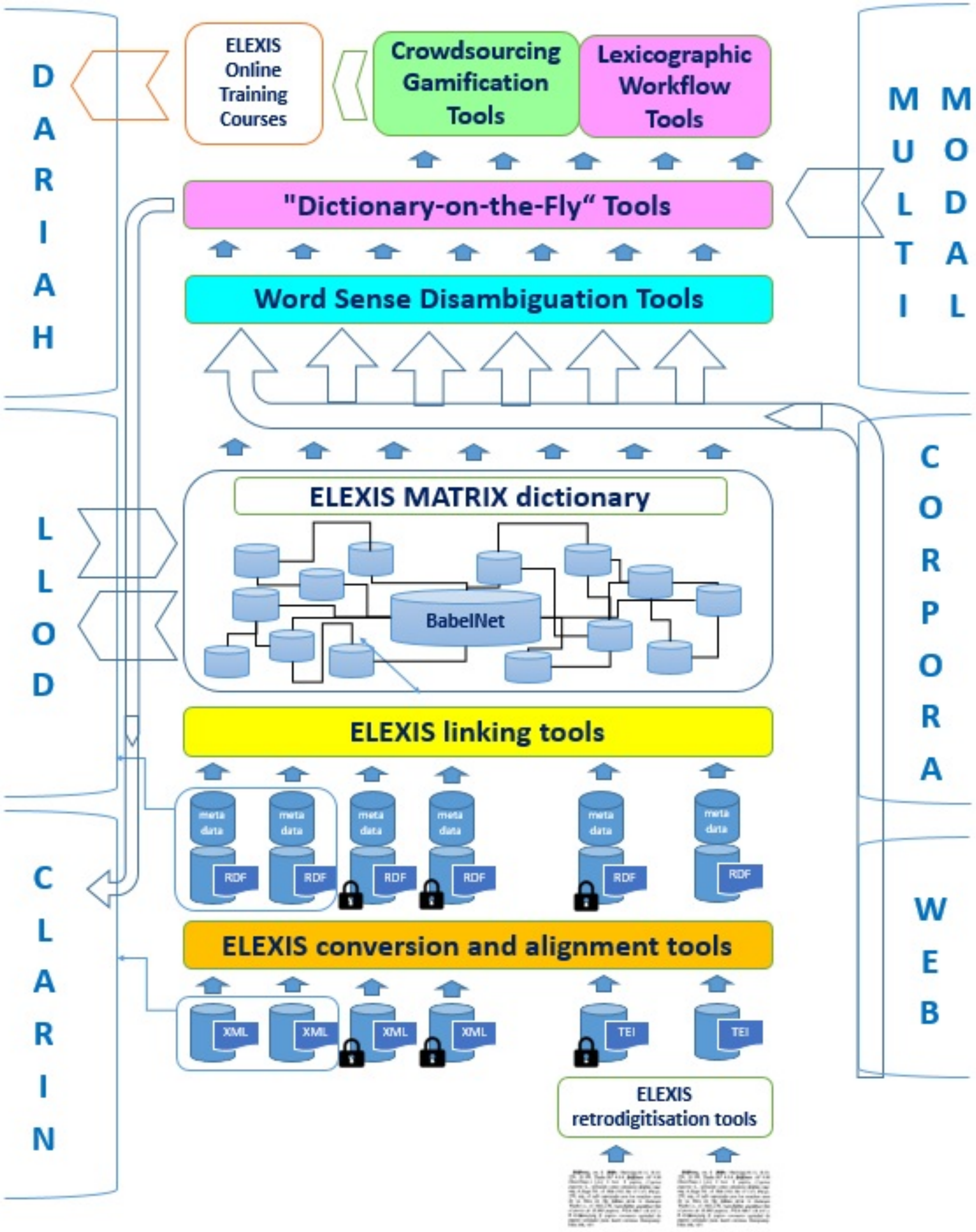
ELEXIS linking tools (direct, via BabelNet, Interlingual Index ...)

ELEXIS conversion and alignment tools (common data models)

ELEXIS retrodigitisation tools

VIRTUAL ACCESS

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


DARIAH
ELEXIS Online Training Courses
Crowdsourcing Gamification Tools
Lexicographic Workflow Tools
MULTI MODAL
"Dictionary-on-the-Fly" Tools
Word Sense Disambiguation Tools
LLOD
ELEXIS MATRIX dictionary
BabelNet
CORPORA
ELEXIS linking tools
meta data
RDF
ELEXIS conversion and alignment tools
XML
TEI
CLARIN
WEB
ELEXIS retrodigitisation tools

VIRTUAL ACCESS

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

What happens with your dictionary?

- (It gets (retro)digitised)
- You are using your own data model for your dictionary (Word, XML etc)
  - CONVERSION TO COMMON DATA MODEL (e.g. TEI, TEI-LEX-0)
- The human oriented common data model is not really NLP friendly
  - CONVERSION TO MACHINE READABLE FORMAT (e.g. Ontolex, Lemon)
- Now it can be linked with other dictionaries via shared concepts
  - LINKING DIRECTLY OR VIA BABELNET
- Shared concepts end up in MATRIX dictionary together with links to your original dictionary
  - MATRIX DICTIONARY, OPEN ACCESS RESOURCE WITH LINKED DATA

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

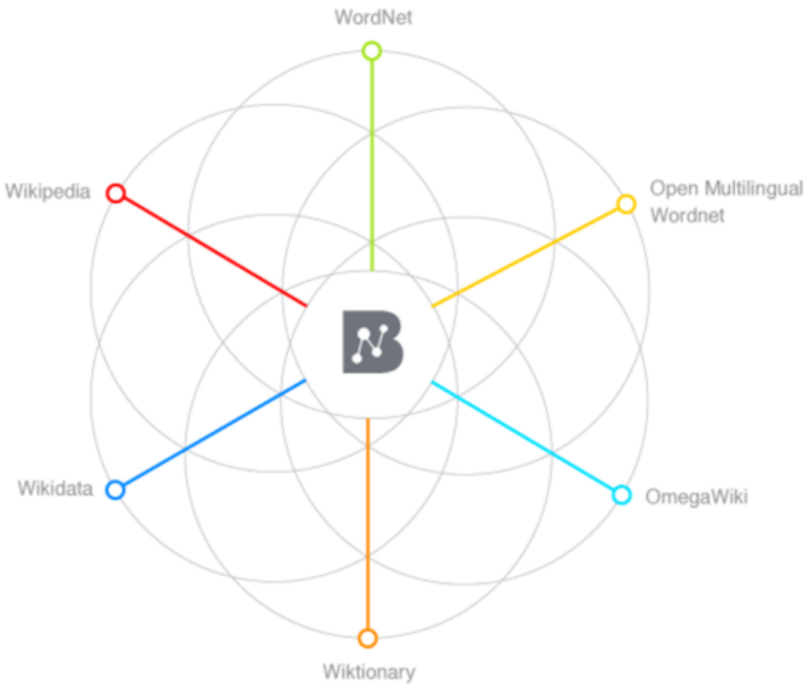
What happens with your corpus?

- It is processed on lower levels (POS-tagging, parsing etc.) and babelfied or wikified
  - WORD SENSE DISAMBIGUATION BASED ON BABELNET OR WIKIPEDIA
- now some methods can be applied to explore semantic behaviour of the vocabulary (also MWEs), to look for translation equivalents etc.
  - SEMANTIC ANALYTICS AND MULTILINGUAL SEMANTIC PARSING
- also, the corpus and the dictionary start functioning almost as one resource in push/pull model
  - ENRICHMENT OF LEXICOGRAPHIC RESOURCES
- the automatically extracted or enriched data can be manually curated in various dictionary writing systems, crowdsourcing platforms or in games
  - DWS, CROWDSOURCING TOOLS, GAMIFICATION

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

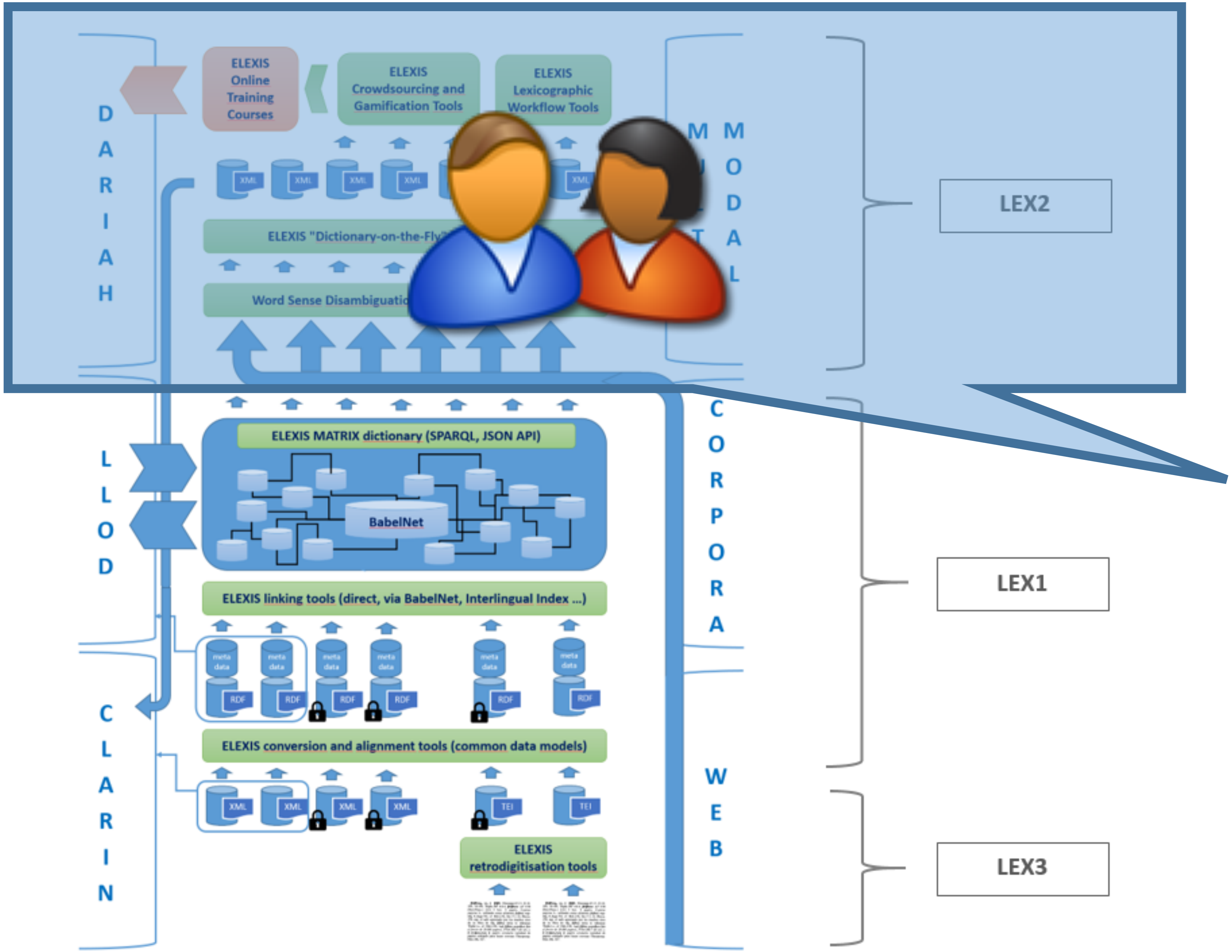
What is in the middle?

- 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.
- a universal **lexicographic metastructure**; a **dictionary matrix** spanning across languages and time



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



DARIAH

LLOD

CLARIN

MULTIMODAL

CORPORA

WEB

ELEXIS Online Training Courses

ELEXIS Crowdsourcing and Gamification Tools

ELEXIS Lexicographic Workflow Tools

XML

ELEXIS "Dictionary-on-the-Fly"

Word Sense Disambiguatio

ELEXIS MATRIX dictionary (SPARQL, JSON API)

BabelNet

ELEXIS linking tools (direct, via BabelNet, Interlingual Index ...)

meta data

RDF

ELEXIS conversion and alignment tools (common data models)

XML

TEI

ELEXIS retrodigitisation tools

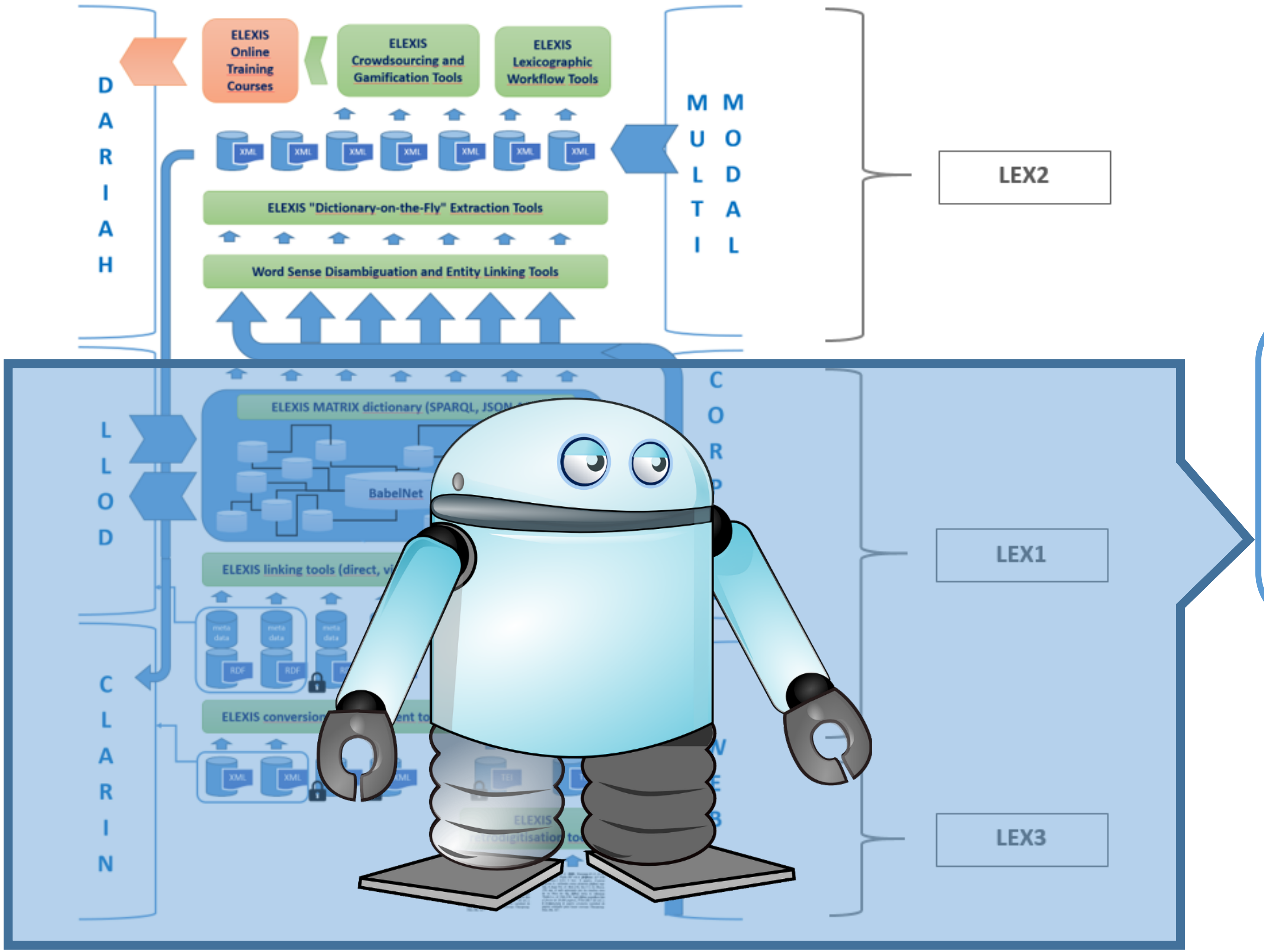
LEX2

LEX1

LEX3

VIRTUAL ACCESS

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


ELEXIS Online Training Courses
ELEXIS Crowdsourcing and Gamification Tools
ELEXIS Lexicographic Workflow Tools
XML
ELEXIS "Dictionary-on-the-Fly" Extraction Tools
Word Sense Disambiguation and Entity Linking Tools
DARIAH
MULTI MODAL
LEX2
ELEXIS MATRIX dictionary (SPARQL, JSON
BabelNet
LLOD
ELEXIS linking tools (direct, vi
meta data
RDF
ELEXIS conversion
ent to
XML
TEI
ELEXIS retrodigitisation to
CLARIN
CORP
WEB
LEX1
LEX3

VIRTUAL ACCESS

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

Presentations

LEX 2 (Session 1: 13.00-14.30)

- **Miloš Jakubíček**
  - Lexical Computing
- **Michal Měchura**
  - Lexical Computing
- **Iztok Kosem**
  - Jožef Stefan Institute

LEX 1/3 (Session 2: 15.00-16.30)

- **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.

Session 3: 17.00–18.00

- Booth 1:
  - General information point
    - Simon Krek, Iztok Kosem, Ondrej Matuska, Tanja Wissik, Anna Woldrich
- Booth 2:
  - Lexicographic data and workflow (Carole Tiberius)
- Booth 3:
  - Interoperability & Linked (Open) Data (John McCrae)
- Booth 4:
  - Lexicographic Data for Natural Language Processing (Roberto Navigli)
- Booth 5:
  - Natural Language Processing for Lexicography (Miloš Jakubíček)
- Booth 6:
  - Lexonomy and GROBID-dictionaries (Michal Mechura, Mohamed Khemakhem)

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

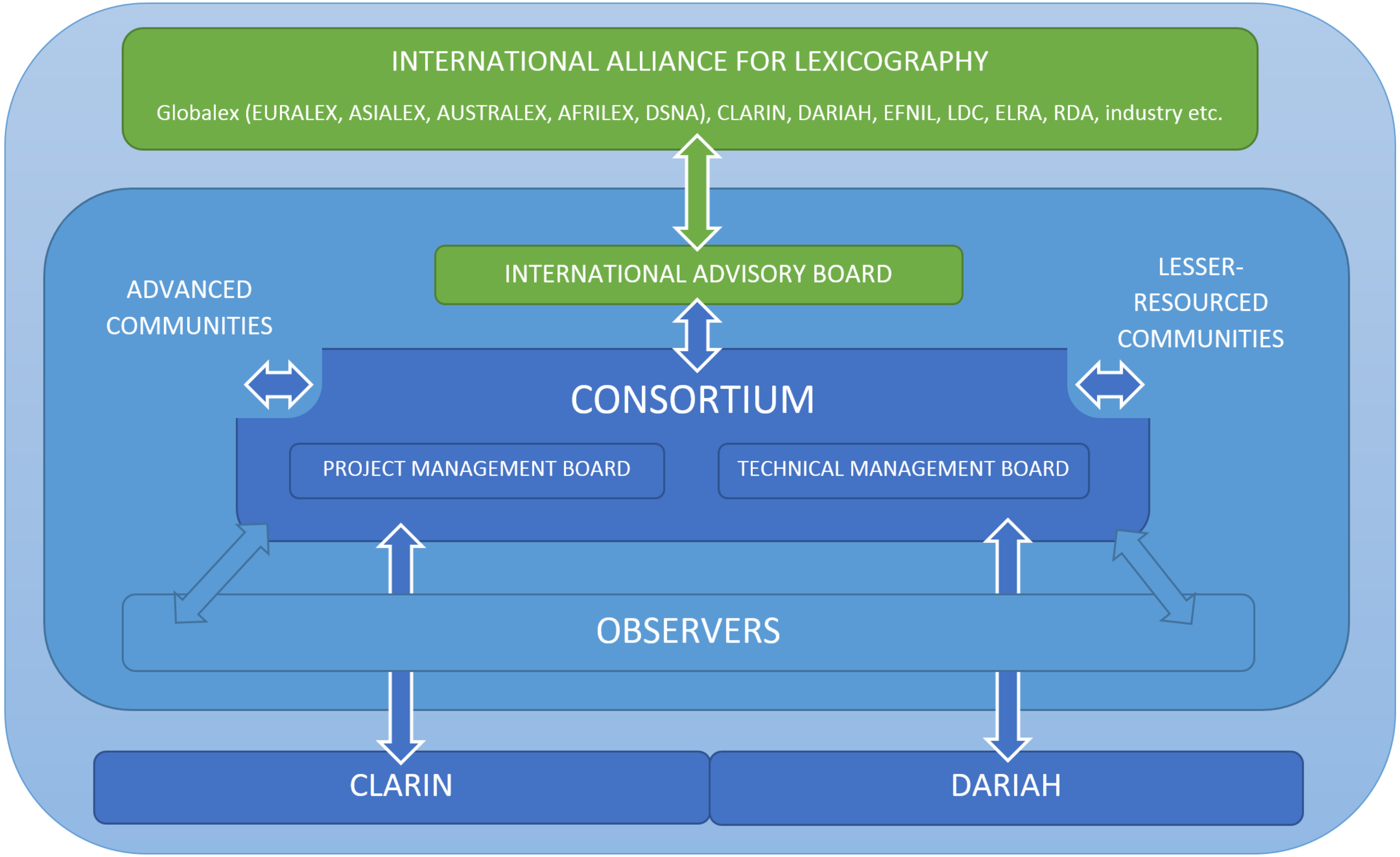
TRANS-NATIONAL ACCESS

- regular calls for **visiting grants** with an average duration of two weeks for researchers to experiment with lexicographical data
- 5 calls for visiting grants will be launched during the project period amounting to an overall number of **30-40 grants**
  - the **calls** will include descriptions of the particular lexicographical resources, tools, and expertise that are made available
  - interested researchers/lexicographers should make an **application**, describing background, purpose, etc., which will be assessed by a committee
  - during the grant visits, the hosting institutions will provide **support** in terms of lexicographical and IT manpower expertise
- Grant winners: poster session (Session 3: Monday, 17:00-18:00)
- ELEXIS travel grants (Session 4: Tuesday, 9.00–10.30)

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



INTERNATIONAL ALLIANCE FOR LEXICOGRAPHY

Globalex (EURALEX, ASIALEX, AUSTRALEX, AFRILEX, DSNA), CLARIN, DARIAH, EFNIL, LDC, ELRA, RDA, industry etc.

INTERNATIONAL ADVISORY BOARD

ADVANCED COMMUNITIES

LESSER-RESOURCED COMMUNITIES

CONSORTIUM

PROJECT MANAGEMENT BOARD

TECHNICAL MANAGEMENT BOARD

OBSERVERS

CLARIN

DARIAH

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

Tuesday - community building day

- **Networking activities**
  - documentation, guidelines, collections of best practices
  - online training modules
  - data seal of compliance for lexicographic data
  - workshops, seminars, and conferences
  - international forum

- Session 4: Tuesday, 9:00-10:30
  - Lexicographers' needs
  - Standards in lexicography
  - Accessibility of scientific production in lexicography
    - ELEXIS Event Registry
    - LexBib proposal
  - International Alliance for Lexicography

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

Sessions 5 and 6: Observers, observers!

- Session 5 (Tuesday, 11.00–12.30)
  - Role of observers
  - Observers and IPR issues (copyright)
  - ELEXIS <-> CLARIN <-> DARIAH
- Session 6 (13.30–15.00)
  - General discussion

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

Questions?

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