ELEXIS project overview

Simon Krek

Coordinator

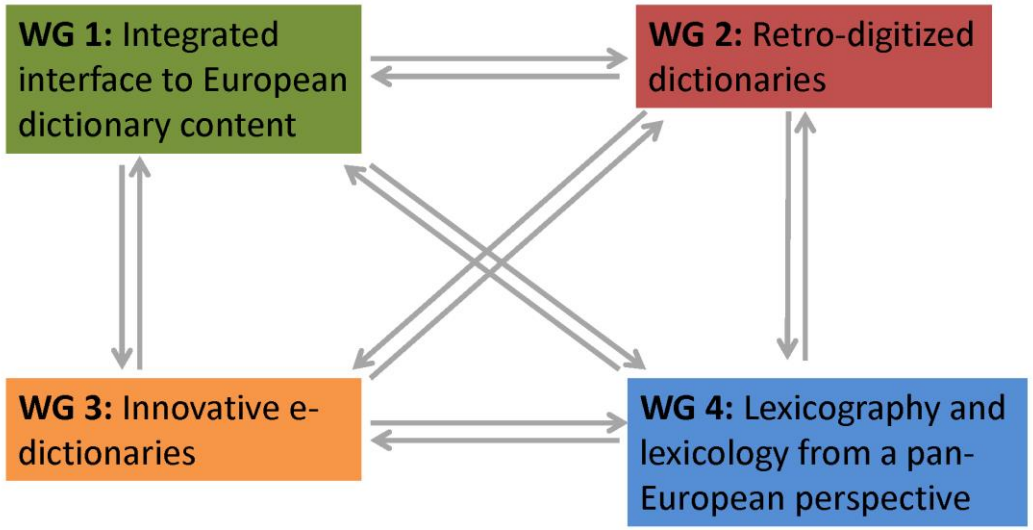
Jozef Stefan Institute

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

COST: European Network of e-Lexicography (ENeL)

October 2013-2017

Working Groups



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.

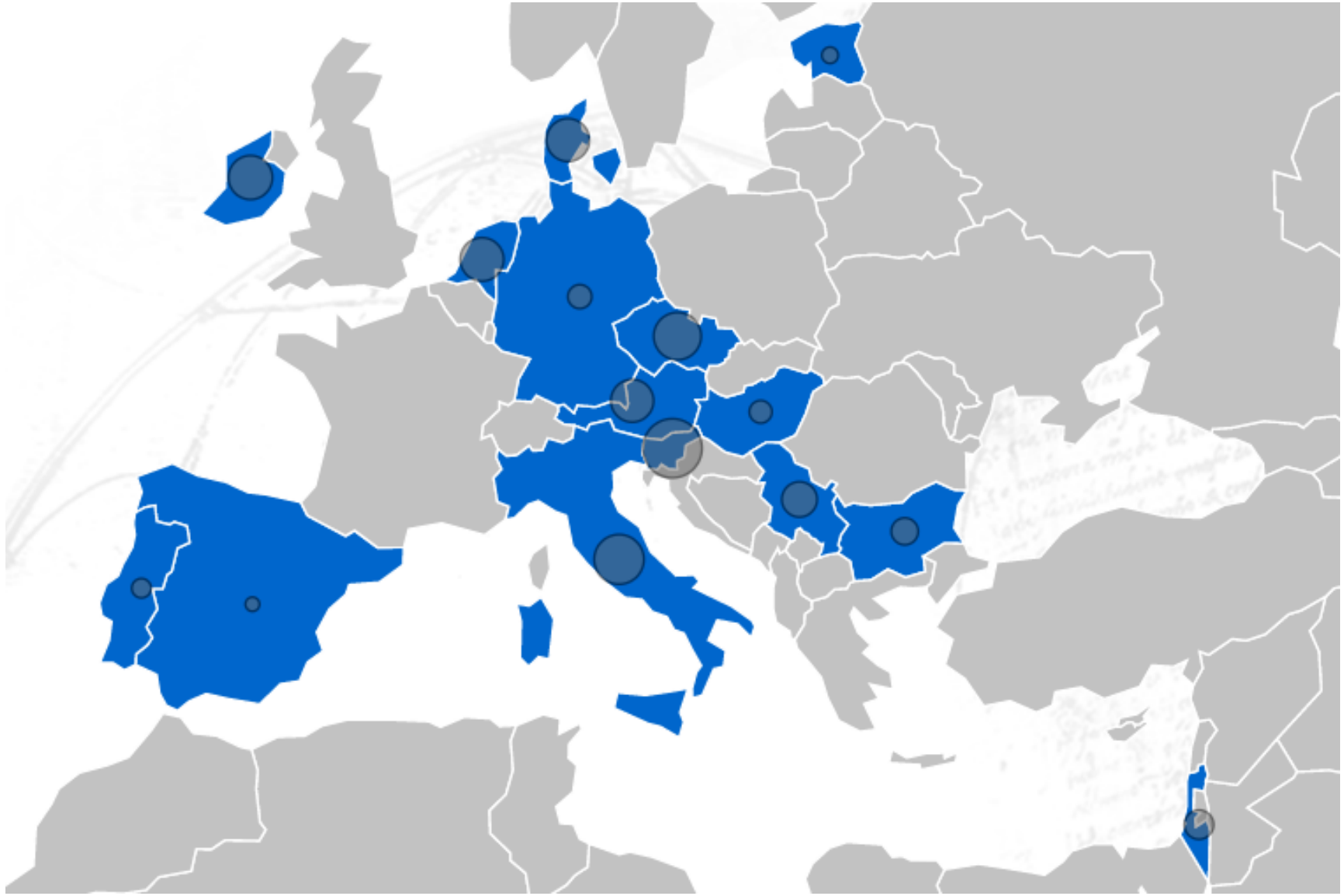
elexis european lexicographic infrastructure

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: 5M €
- 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.

elexis european lexicographic infrastructure



- 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

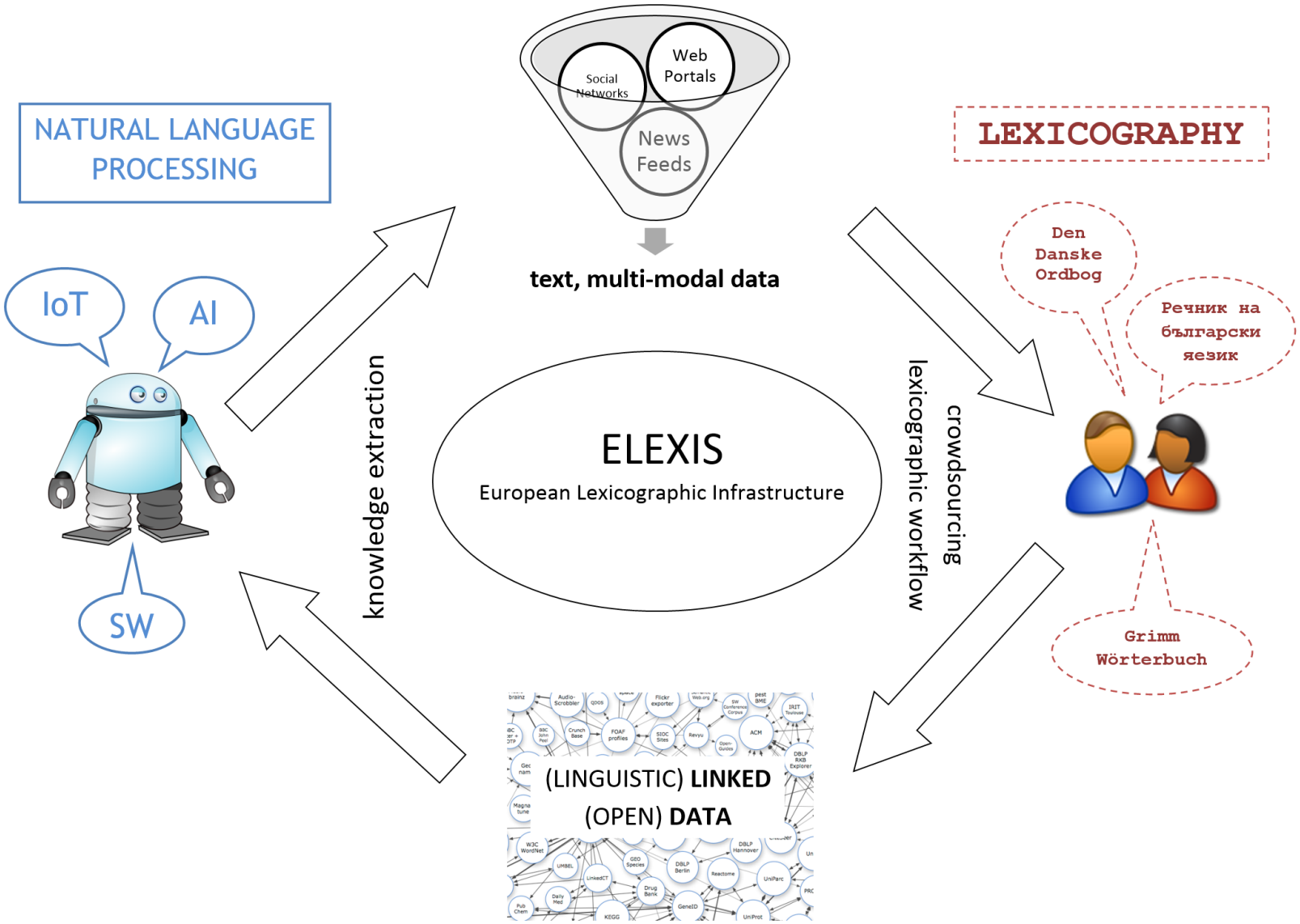
- Providing efficient **access to** quality lexicographic **data**
- Enabling massive **integration of** knowledge-based **resources**
    - Facilitating inclusion of innovative lexicography in **research** and **education**
    - Enabling the use of new technology and data in **industry**
    - Establishing **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.

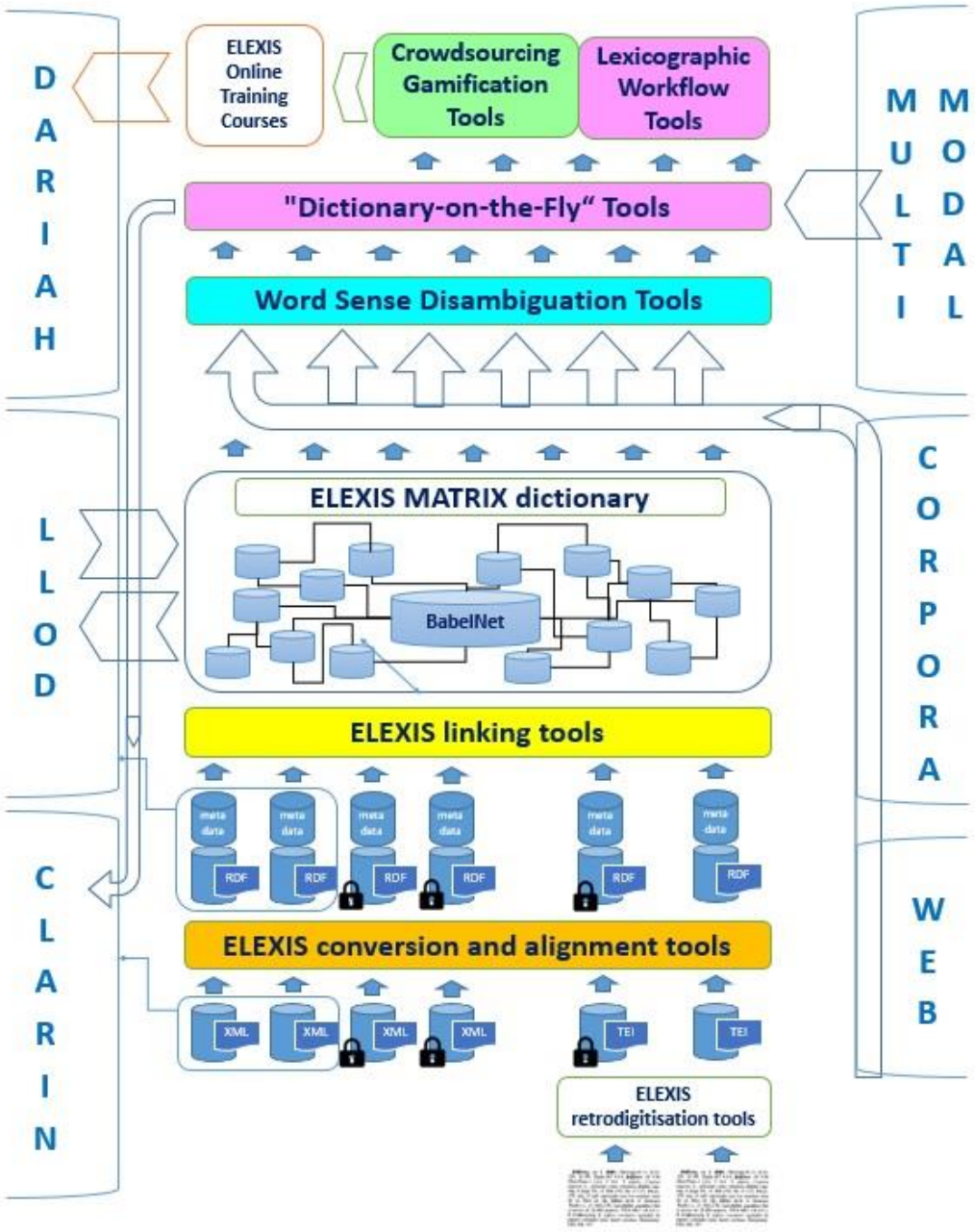


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

THREE TYPES OF ACTIVITIES

- Joint **research** activities
  - tools and methods for enabling linked lexicographic resources
  - 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

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



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.

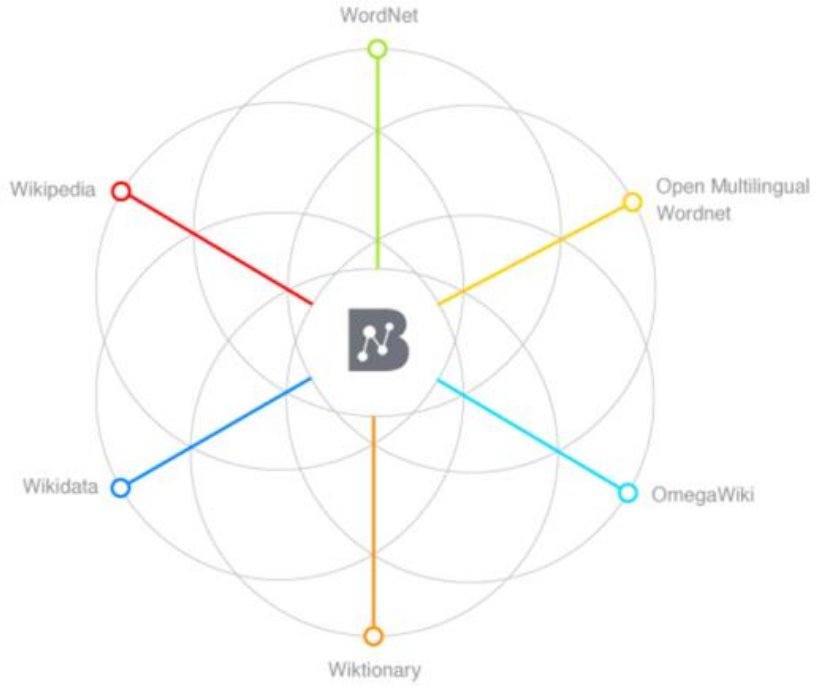
What happens with your corpus?

- 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 meta-structure**; a **dictionary matrix** spanning across languages and time

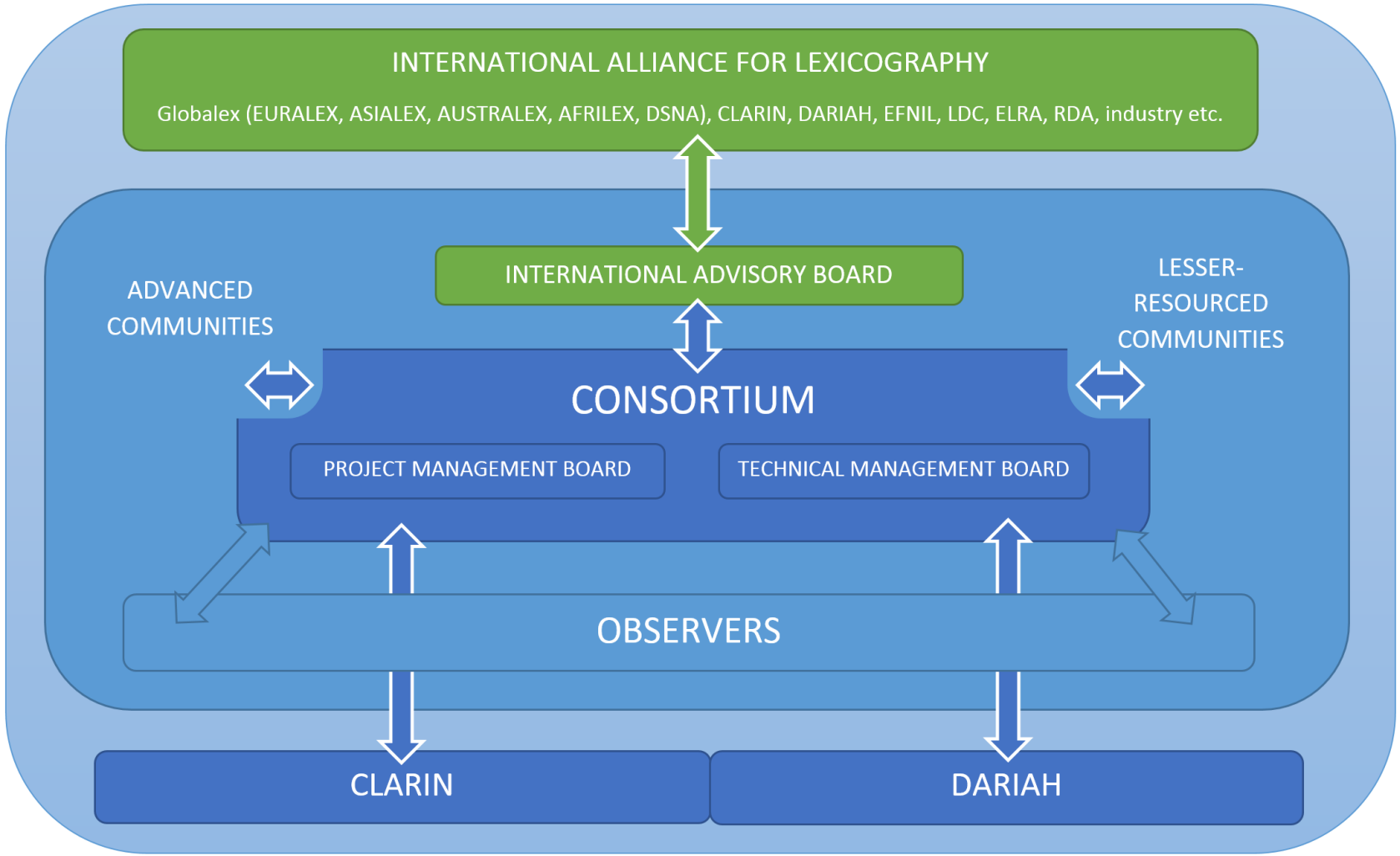


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

elexis european lexicographic infrastructure

ELEXIS

European Lexicographic Infrastructure



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

Thank you

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