Dictionary Matrix

Simon Krek (Slovenia)

Jožef Stefan Institute (Artificial Intelligence Laboratory)

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

Dictionary Matrix: from the Grant Agreement

- ELEXIS aims to develop an infrastructure which will:
- Objective 3
  - develop strategies, **tools** and standards for extracting, structuring and **linking** of lexicographic resources

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

Dictionary Matrix: from the Grant Agreement

- Extensive linking of existing lexicographic resources by pivoting through BabelNet will enable the creation of what we call ELEXIS dictionary matrix –
- a 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 (=Lexonomy).

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

No linking without (open) dictionary data

- Partner data
- Observer data
- Open data
  - Including automatically created dictionaries

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

IPR issues deliverable (web page)

- Licensing and Intellectual Property rights issues connected to data or software should be carefully considered at the very beginning of a lexicographic project, i.e. at the planning or proposal writing stage.
- If an open license cannot be used for the entire lexicographic dataset, using different licenses for different types/parts of lexicographic data should be considered.
- IPR limitations for existing lexicographic resources were consideres as one of the main risks in the project proposal
- (Development of Slovene in Digital Environment experience)

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

Partner data

- 51 resources mentioned in grant agreement
- 32 contributed as specified
- REPLACEMENTS
  - 7 Estonian resources replaced by EKI Combined
  - 1 Hungarian resource
  - 3 OEAW resources
- Licenses
  - Restricted: 22
  - Academic: 19
  - Unknown: 7
  - Public: 3

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

Partner data (1)

JSI	Slovene Lexical Database	Slovene	PUB
INT	Dictionary of Contemporary Dutch (ANW)	Dutch	RES
INT	Dictionary of the Dutch Language (WNT)	Dutch	RES
INT	Dictionary of Old Dutch (ONW)	Old Dutch	RES
INT	Dictionary of Early Middle Dutch (VMNW)	Early Middle Dutch	RES
INT	Dictionary of Middle Dutch (MNW)	Middle Dutch	RES
OEAW	Dictionary of Bavarian Dialects of Austria	Austrian	UNK
OEAW	Dagaare-Cantonese-English Dictionary	Dagaare, Cantonese, English	PUB
OEAW	Haussa-English Dictionary	Haussa, English	UNK
OEAW	Database of Bavarian Dialects of Austria	Austrian Variants	UNK

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

Partner data (2)

OEAW	Russian Dialect Dictionary	Russian	UNK
OEAW	Tunico	Tunisian	PUB
OEAW	Viennese Historical Dictionaries Online (ViDi)	Austrian Variants	UNK
BCHD	Karadžić, Serbian Dictionary (1818, 1852)	Serbian	ACA
BCHD	Mikloshic, Lexicon Palaeoslovenico-Graeco-Latinum (1862—1865)	Old Church Slavic	ACA
BCHD	Daničić, Dictionary of Serbian Literary Antiquity (1863-4)	Serbian (medieval)	ACA
BCHD	Bojanić & Trivunac, Dictionary of Dubrovnik Dialect	Serbian (dialect)	ACA
BCHD	Elezović, Dictionary of Kosovo-Metohija Dialect	Serbian (dialect)	ACA
BCHD	Zlatanović, Dictionary of Southern Serbian Dialects	Serbian (dialect)	ACA
BCHD	Žugić, Dictionary of Jablanica Region	Serbian (dialect)	ACA

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

Observer data status (provided)

mentioned/promised	**124**
provided	71
not yet provided	46
knowledge	8

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

Observer data status

type of data	provided	not provided	All
lexicographic resource	**39**	**19**	**58**
word list	4	0	4
terminology	14	4	18
corpus	9	3	12
knowledge	8		
other	19	5	24
ALL	**71**	**45**	**124**

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

Observer data status (licenses)

License	Lexicographic resources
Creative Commons	**24**
GNU General Public license	**1**
RES	2
Unknown	12
all	39

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

Summary

- IPR issues and availability continue to be a BIG problem
- My expectations:
  - With more AI, more lexicographic data will be available (but not yet)
  - Fkjd
- The tools are here, waiting for the data ☺

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

Demo

- Gregor Leban (ER/IJS)

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