Elexis Observer Event, Vienna, Feb. 18-19, 2019

The LexBib Project

A Corpus and Bibliography of Metalexicography

David Lindemann

Universität Hildesheim

UPV/EHU University of the Basque Country

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

LexBib Project

- Goals
- Features, Methods, Workflow
- Partners
- Working plan, Outlook

- Detailed intro: Lindemann, Kliche & Heid 2018 (Euralex Ljubljana)

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

State of the Art and Goals

Situation now

- Bibliographies: Unlinked islands
  - Publishers / topics
- Noisy metadata sets
  - Automatically harvested
- Language barrier
  - Mutual invisibility
- Unconsistent indexation
  - Keywords / subject headings

LexBib Goals

- Unified territory
  - But: only relevant content
- Complete and correct metadata
  - manually validated
- Multilingual access structure
  - Multilingual Domain Ontology
  - Additional content-describing metadata

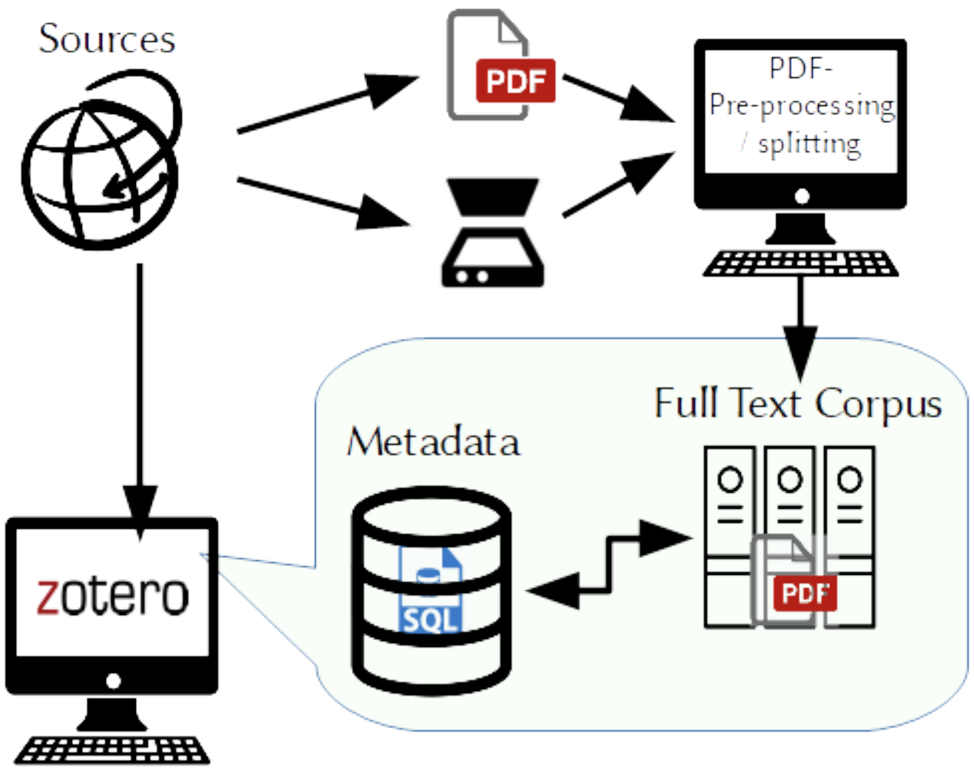
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

LexBib online reference platform

- Advanced search & browsing functions
- Access through domain ontology nodes
- Access through citation network
  - Visualisation of network and clusters
- Author profiles
  - Author name disambiguation, additional info
- Link to full text
  - …on publishers’ etc. platforms

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

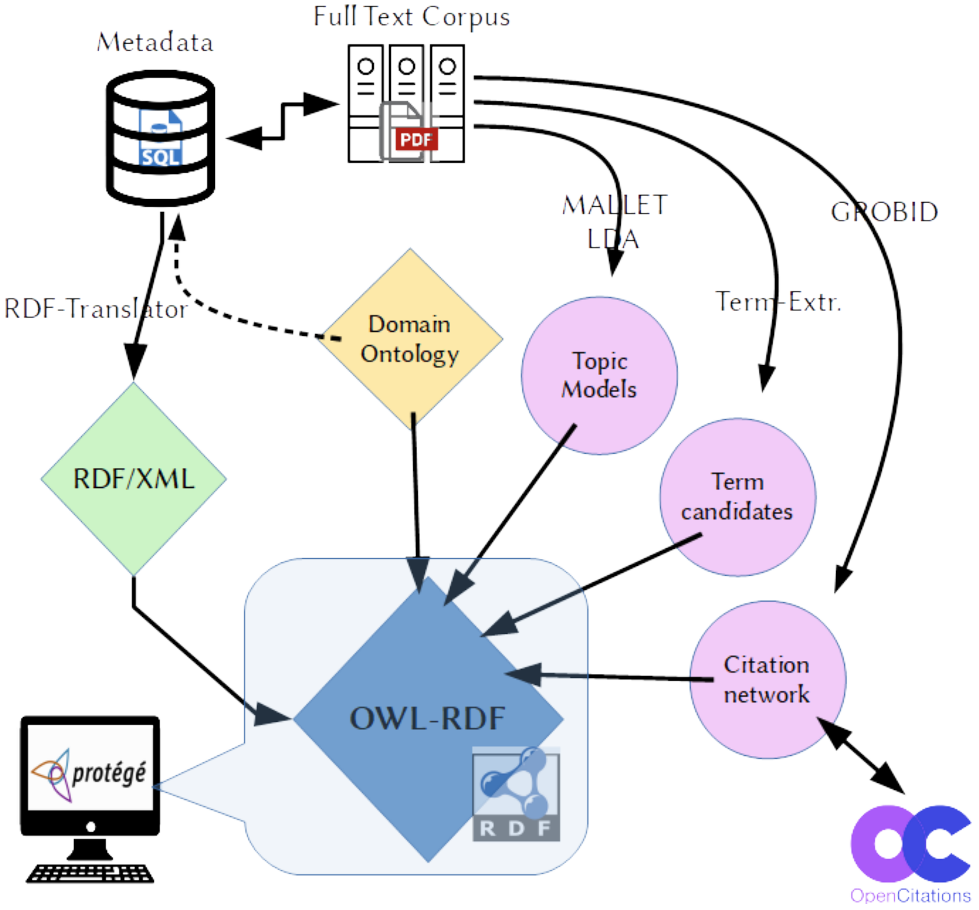
Working Pack 1



- Collection and validation of Metadata
  - If necessary
    - completion/correction
- Collection of full texts
  - If necessary
    - Scan
    - OCR
    - Split (collective volumes)

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

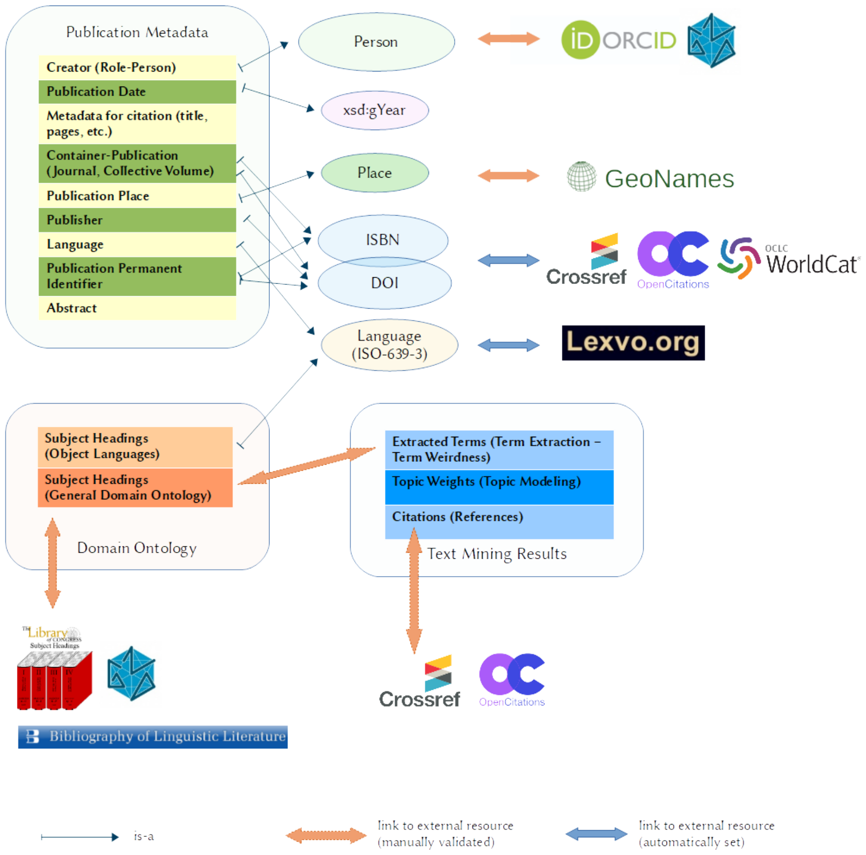
Working Pack 2



- Extraction from full texts
  - Term / keyword candidates
    - Weirdness / reference corpora
    - Tools: Rösiger, Heid et al. (IMS Stuttgart)
  - Topic Models
    - Topic Weights (for recommender)
    - Tools: McCallum: MALLET
  - Citation relations
    - Citation network, clusters
    - Tools: Romary, Lopez et al. (GROBID)

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

Working Pack 3


Publication Metadata

Creator (Role-Person)

Publication Date

Metadata for citation (title, pages, etc.)

Container-Publication (Journal, Collective Volume)

Publication Place

Publisher

Language

Publication Permanent Identifier

Abstract

Person

xsd:gYear

Place

ISBN

DOI

Language (ISO-639-3)

Subject Headings (Object Languages)

Subject Headings (General Domain Ontology)

Domain Ontology

Extracted Terms (Term Extraction – Term Weirdness)

Topic Weights (Topic Modeling)

Citations (References)

Text Mining Results

Bibliography of Linguistic Literature

is-a

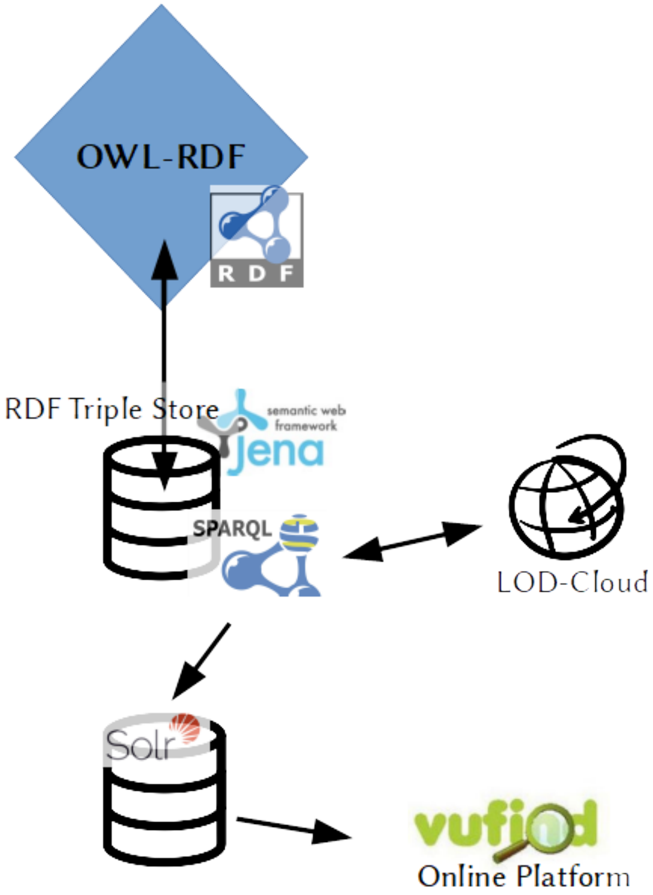
link to external resource (manually validated)

link to external resource (automatically set)

- Domain Ontology
  - Controlled Vocabulary of subject headings
    - Root node “Lexicography”, linked to LCSH, DNB, Wikidata…
  - Multilingual labels
    - EN, DE, ES, …
  - Sources: Existing classifications
- Linked Open LexBib
  - Use of existing element sets
  - Links to existing resources

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

Working Pack 4



- Open Zotero Group
  - zotero.org/groups/LexBib
- Linked Open LexBib
  - SPARQL Endpoint
- LexBib Online Platform
  - Implementation with *vufind*

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

Working Pack 5

- Dissemination, Community
  - Euralex, eLex conferences
    - Feedback
    - Call for contributions
    - Curation of author pages

- Process Metadata
  - Predictions for similar projects on a larger scale:

  “How much manual work is necessary for a comparatively noise-free and complete resource?”

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

Partners

DFG-Proposal (in prep.): Partners

- Uni Hildesheim: Lead, WP 1-5
- BBAW: WP 2, 5
- IDS: WP 1, 4, 5

Financial and other support

- Uni Hildesheim: Preliminaries
- Elexis: Preliminary work WP 1, 3
- Euralex
- EMLex
- Lexicografía UDC
- LOC-DB / VZG
- De Gruyter

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

Working Plan

- 2019 [Uni Hildesheim, Elexis]
  - Metadata: Euralex-Dykstra, Obelex-Meta, Wiegand, Córdoba, Ahumada
  - Full Texts: English 2000-2018
  - Domain Ontology: First proposal
- 2020-22 [DFG?]
  - Metadata: All European languages, 1971-2018
  - Full texts: German, Spanish 2000-2018 (and more?)
  - WP2-5
- In a near future
  - Other languages, older full texts

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

Thank you!
Questions?

david.lindemann@uni-hildesheim.de

http://zotero.org/groups/LexBib

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