



Bibliodata LOD-ification in the **LexBib** project: Author disambiguation



David Lindemann david@lexbib.org
Christiane Klaes christiane@lexbib.org



european lexicographic
infrastructure



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

Introduction

Project: "LexBib" Digital Bibliography Platform of Lexicography and Dictionary Research

Items as instances of domain ontology classes

Publications

- publication metadata, structured
- content describing metadata
 - subject indexation, citation networks, extracted terms, topic models, ...

Persons (article authors, book editors, resource creators)

Organisations (linked to resources, projects, persons, places)

Projects (linked to persons, resources, articles)

Events (conferences, workshops, linked to articles, places, organisations)

Places (linked to events, persons, organisations, projects, events)

Workflows

Structuring of plain text bibliographies (stand-alone and in-article bibliographies)

LOD-ification (URI for entities, and relations defined in LOD vocabulary)

DARIAH WG Bibliodata as platform for transparent and collaborative workflow development

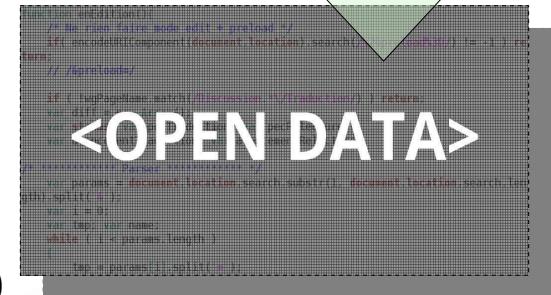
Semantic Web standards, free and open source tools

21923 Voillat, François: Le "Glossaire des patois de la Suisse romande" (GPSR). In: Actes du XVIII^e Congrès International de Linguistique et de Philologie Romanes [...]. Tome VII, 1989↑, 338–345.

15475 Nyhlén, Lars-Olaf: Wie sagt man in Österreich? In: Moderna Språk 68. 1974, 275–281. [Bes. zu: Jakob Ebner: Duden. Wie sagt man in Österreich? Wörterbuch der österreichischen Besonderheiten. Mannheim 1969 (Duden Taschenbücher 8)].

15397 Norden 11.–14. maj 1993. Red. af Anna Garde og Pia Jarvad. Oslo 1993 (Skinner udgivet af Nordisk Forening for Leksikografi. Skrift nr. 2). [Daraus: 363, 1003, 2755, 3101, 3870, 5222, 5251, 5390, 6058, 6758, 9206, 9743, 10158, 11521, 11553, 11843, 12600, 12940, 12974, 13146, 13399, 14192, 15316, 15345, 15402, 15474, 15651, 15944, 17322, 20818, 21371].

15427 Novikov, L. A.: K probleme omonimii. [Zum Problem der Homonymie]. In: Leksikografičeskij sbornik 1960, 93–102.



LexDo Domain Ontology

- ▼ Documentation: <http://lexbib.org/lexdo>
- ▼ Standard RDF vocabularies
 - ▼ BIBO, PROV, Dublin Core, FOAF, Ontolex
- ▼ Top-level concepts:
 - ▼ Agent
 - ▼ Person/Organization/Software
 - ▼ Resource
 - ▼ Bibliographical resource/Lexical resource/Software
 - ▼ Subject
 - ▼ Subjects of metalexicographical items
 - ▼ Language
 - ▼ Event
 - ▼ Place

Ontology Specification Draft

LexDo - Domain Ontology of Lexicogra

Release 2020-03-14T16:50:24.513+01:00

This version:

<http://lexbib.org/lexdo/1.0.2/>

Authors:

[David Lindemann](#)

Contributors:

[Christiane Klaes](#)

[Laura Giacomini](#)

Publisher:

Institut Jožef Stefan / Universität Hildesheim / UPV/EHU Uni

Download serialization:

Format [JSON LD](#) Format [RDF/XML](#) Format [N Triples](#) Format [TTL](#)

License:

License <http://creativecommons.org/licenses/by/3.0/> License

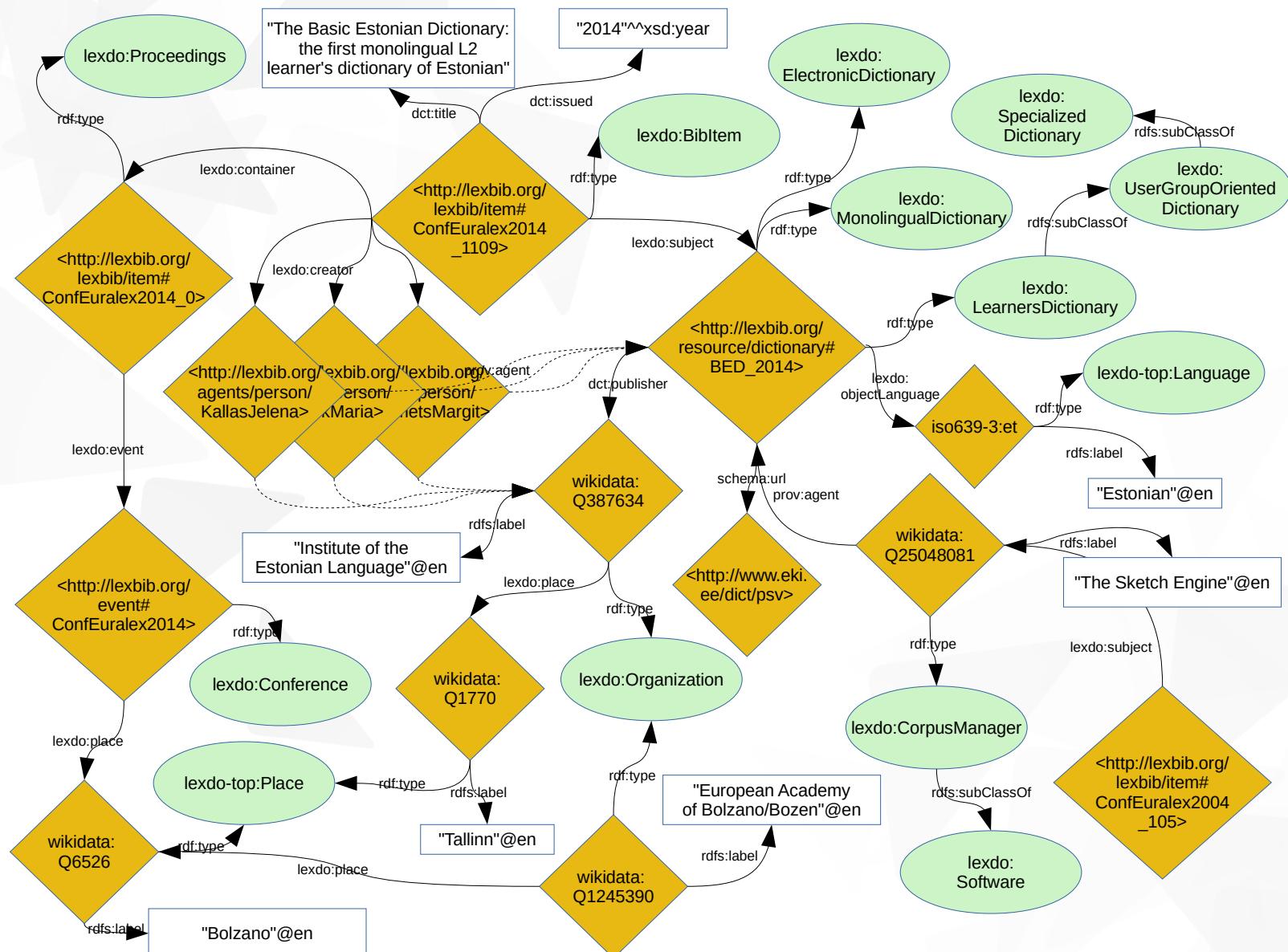
Cite as:

LexDo Domain Ontology of Lexicography and Dictionary Res
For full reference, see [About LexBib](#) at Zotero.

<http://lexbib.org/lexdo/1.0.2>



LexDo: Towards Lexicography as Knowledge Graph



LexBib Zotero Collection

zotero

The screenshot shows the Zotero web interface. On the left, a sidebar titled "Group Libraries" lists several groups under "LexBib": ABOUT_LEXBIB, BIBLIOGRAPHIES, COLLECTIVE_VOLS, CONFERENCES, Euralex_2018_experiment (which is selected), JOURNALS, MONOGRAPHS, and OBELEX-Meta-Part. The main area displays a list of academic papers in a table format. The columns are "Creator", "Date", and "Title". The first few titles are: "A Call for a Corpus-Based Sign Language Dictionary: An ...", "A Good Match: a Dutch Collocation, Idiom and Pattern Di...", "A lexicon of Albanian for natural language processing", "A Sample French-Serbian Dictionary Entry based on the ...", "A Universal Classification of Lexical Categories and Gram...", "A Workflow for Supplementing a Latvian-English Dictiona...", "Advances in Synchronized XML-MediaWiki Dictionary De...", "An Overview of FieldWorks and Related Programs for Co...", and "Analyzing User Behavior with Matomo¹ in the Online Info...". Below the table, there are tabs for "Info", "Notes", "Tags", "Attachments", and "Related". Under the "Info" tab, detailed metadata is shown for the selected paper: Item Type (Conference Paper), Title (A Sample French-Serbian Dictionary Entry based on the ParCoLab Parallel Corpus), Author (Marjanović, Saša; Stosic, Dejan; Miletic, Aleksandra), Editor (Čibej, Jaka; Gorjanc, Vojko; Kosem, Iztok; Krek, Simon), and Abstract (a text about the paper's aim and methodology). A search bar at the top right contains the placeholder "Title, Creator, Year".

Access:

- ▶ either: <http://lexbib.org/zotero>
- ▶ or: <https://www.zotero.org/groups/lexbib>
- ▶ All metadata present, searchable
- ▶ structured metadata for local citation managers, export options
- ▶ Collaborative editing possible
- ▶ Status: Check here



Zotero: all metadata are literal values

The screenshot shows the Zotero desktop application. On the left is a sidebar with categories like LexBib, COLLECTIVE_VOLS, CONFERENCES, JOURNALS, and DSNA. The DSNA folder is selected. The main pane displays a list of bibliographic entries. A context menu is open over the entry for 'Dating Phonological Change on the Basis of Eighteenth-Century British English Dictionaries and Orthoepic Treatises' by Trapateau, Nicolas. The menu shows various metadata fields with their corresponding values.

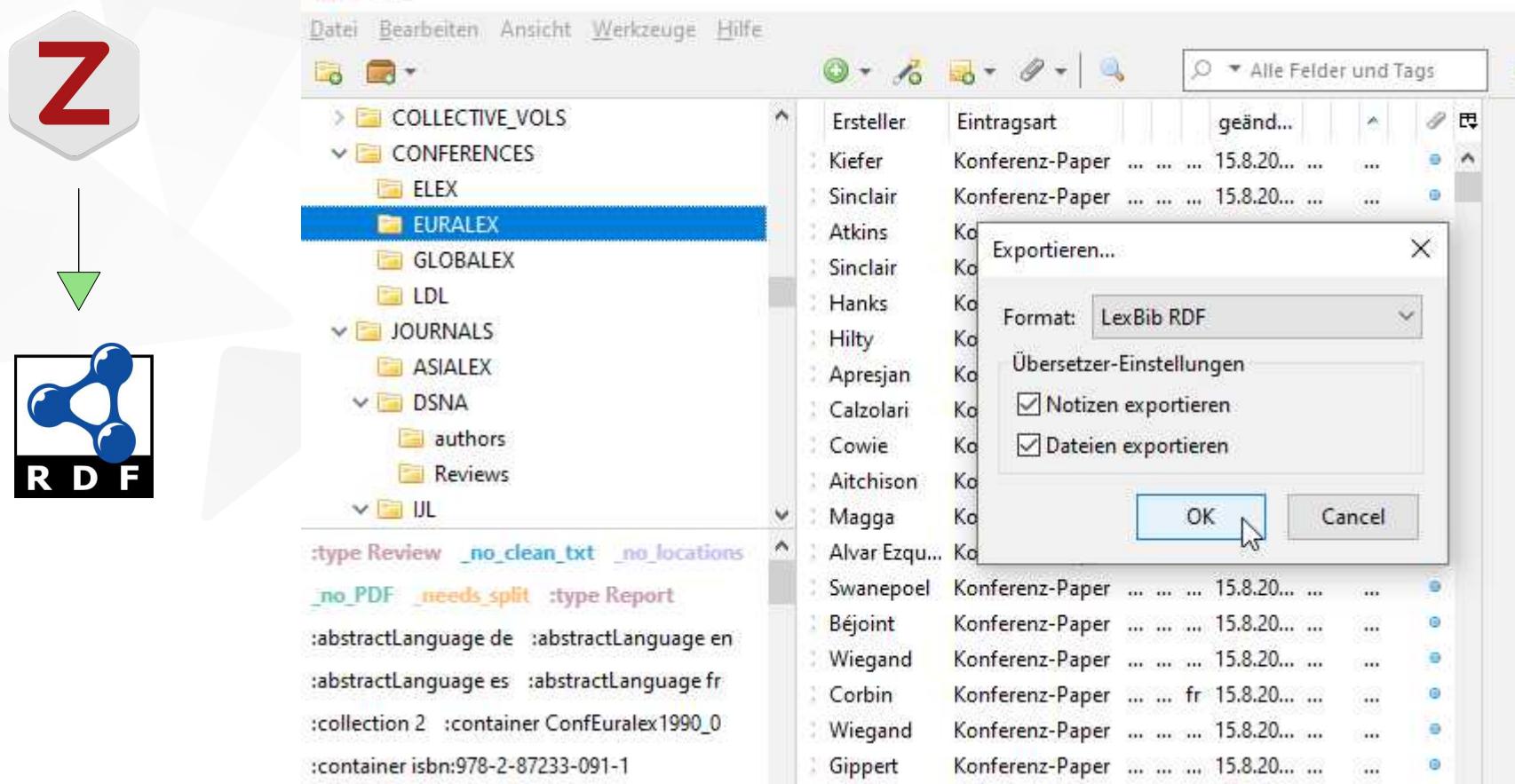
Titel	Ersteller	Jahr	S...
FORUM: Using OE...	Coleman	2013	... 1-9
Frederic Gomes Ca...	Hall	2001	... 1-13
The Revolution of ...	Simpson	2002	... 1-15
Guarding Even Our E...	Knowles	2015	... 1-16
Webster's Third an...	Schütz	2009	... 1-17
A Functional Appr...	Nielsen	2006	... 1-20
European Lexicogr...	Hartmann	2000	... 1-21
"Decent Reticence...	Mugglestone	2007	... 1-22
The Mysterious Ca...	Ogilvie	2008	... 1-22
Birds of a Feather?...	Male und Rodriguez	2003	... 1-27
Dating Phonologic...	Trapateau	2017	... 1-29
Contrasts in America...	Murphy	2018	... 1-30
Shared Lexical Inn...	Peters et al.	2019	... 1-30
James Boswell (17...	Caudle	2011	... 1-32
The Dictionary Soc...	Adams	2014	... 1-35
A New Typology of C...	Williams	2016	... 1-35
"Academic Hoolig...	Farina und Durman	2012	... 1-41
The "Electronificati...	Brewer	2004	... 1-43
Illustrating Webster	Hancher	2010	... 1-45
The Decline of the ...		2017	... 1-46

Eintragsart: Zeitschriftenartikel
Titel: Dating Phonological Change on the Basis of Eighteenth-Century British English Dictionaries and Orthoepic Treatises
Autor: Trapateau, Nicolas
(...) Zusammenfassung: Lexicographic evidence from eighteenth-century English...
Publikation: Dictionaries: Journal of the Dictionary Society of North America
Band: 38
Ausgabe: 2
Seiten: 1-29
Datum: Dezember 14, 2017
Reihe:
Titel der Reihe:
Reihe Text:
Zeitschriften-Abkürzung:
Sprache: en
DOI: 10.1353/dic.2017.0018
ISSN: 2160-5076
Kurztitel: DSNA Journal 38-2 (2017)
URL: http://muse.jhu.edu/article/679602
untergeladen am: 11.5.2020
Archiv:
standort im Archiv:
bibliothekskatalog: Project MUSE
Signatur:
Rechte:
Extra: https://en.wikipedia.org/wiki/Nice

- ▶ For some categories, simple mapping solves it:
 - ▶ Languages > ISO-639 codes on LEXVO, Wikidata
 - ▶ DOI, ISBN, ISSN > on Wikidata, etc.

Zotero > LexBib RDF

- Own script for LexBib RDF export: See Elexifinder at Github



LexDo properties in Zotero

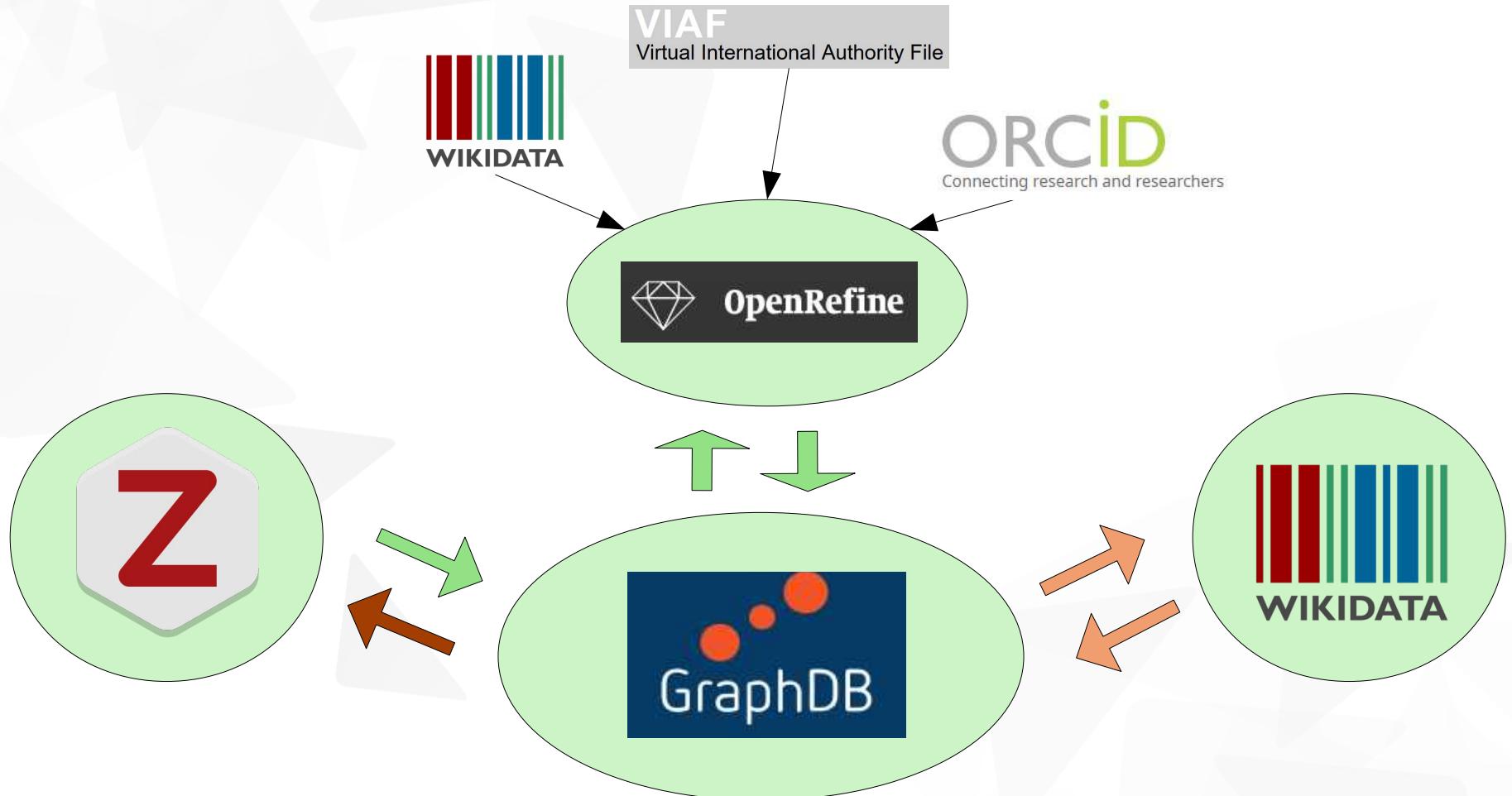
The screenshot shows the Zotero application interface. On the left, there's a sidebar with a tree view of collections, including 'LexBib' which contains various sub-folders like 'ABOUT_LEXBIB', 'COLLECTIVE_VOLS', 'CONFERENCES', 'ELEX', 'EURALEX', 'GLOBALEX', 'LDL', 'JOURNALS', 'ASIALEX', 'DSNA', 'IJL', 'LEXICOGRAPHICA', 'LEXICON', 'LEXICONORDICA', 'LEXIKOS', 'NSL', 'VARIOUS', and 'VIDEOS'. Below the sidebar, there's a list of URLs starting with ':container https://tidsskrift.dk/nsil/issue/view/'. In the main pane, a list of items is shown with columns for 'Titel', 'Ersteller', 'Standort im Archiv', and 'Infos'. One item is selected, highlighted with a blue border. The right panel displays detailed properties for this selected item, such as 'Eintragsart: Zeitschriftenartikel', 'Titel: LEXIA – en islandsk-fransk online ordbog: Udfordringer og løsninger', 'Autor: Daviðsdóttir, Ró...', 'Zusammenfassung: Nordiske Studier i Leksikografi', 'Publikation: Band 14', 'Ausgabe: Seiten 63-70', 'Datum: 2018', 'Reihe: Titel der Reihe', 'Text: Reihe Text', 'Zeitschriften-Abkürzung: Sprache da', 'DOI: ISSN 2246-7823', 'Kurztitel: Nordiske Studier i Leksikografi 14 (2018)', 'URL: https://tidsskrift.dk/nsil/articl...', 'Heruntergeladen am: 28.4.2020, 23:50:47', 'Archiv: Standort im Archiv: https://tidsskrift.dk/nsil/articl...', and 'Bibliothekskatalog: tidsskrift.dk'. A green arrow points from the 'lexdo:publisher' tag in the left panel to the 'lexdo:publisher' property in the right panel.



(lexdo:BibItem instance) lexdo:publisher (Wikidata URI)

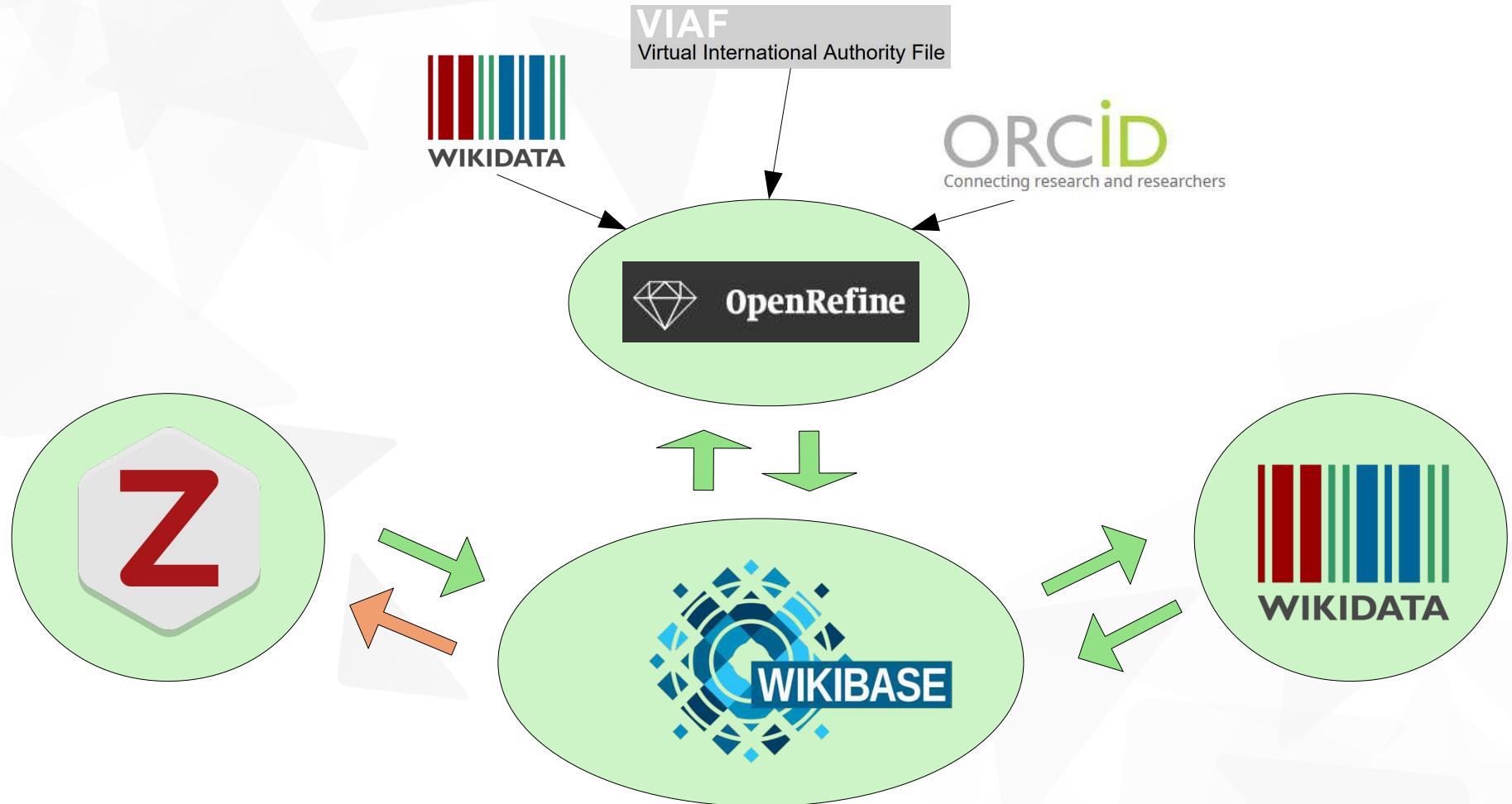
- straightforward
- not trivial but possible
- not possible

Data migration scheme using Triple Store

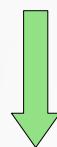


- straightforward
- not trivial but possible
- not possible

Data migration scheme using Wikibase



Zotero > LexBib RDF: Interim Author URI, string based



```
22      <zotexport:pdfFile>D:/Zotero/storage/BXTR4R44/Adamska-SalaciakA_2016_fullte
23      <zotexport:txtFile>D:/Zotero/storage/A7L6MWCY/euralex_2016_084_p758.txt</zot
24  </owl:NamedIndividual>
25  <owl:NamedIndividual rdf:about="https://euralex.org/publications/on-bullying-mo
26    <rdf:type rdf:resource="http://lexbib.org/lexdo/BibItem"/>
27    <lexdo:publishingPlace>Tbilisi</lexdo:publishingPlace>
28    <lexdo:publisherName>Ivane Javakhishvili Tbilisi State University</lexdo:pu
29    <bibo:pages>758-766</bibo:pages>
30    <bibo:isbn13>978-9941-13-542-2</bibo:isbn13>
31    <lexdo:eventName>XVII International EURALEX Congress</lexdo:eventName>
32    <dcterms:title>On Bullying, Mobbing (and Harassment) in English and Polish</d
33    <lexdo:containerTitle>Proceedings of the 17th EURALEX International Congres
34    <lexdo:containerShortTitle>Euralex (2016)</lexdo:containerShortTitle>
35    <lexdo:creator rdf:resource="http://lexbib.org/agents/person/Adamska-Sałac
36    <bibo:authorList rdf:resource="https://euralex.org/publications/on-bullying-mo
37    <lexdo:editor rdf:resource="http://lexbib.org/agents/person/MargalitadzeTim
38    <bibo:editorList rdf:resource="https://euralex.org/publications/on-bullying-mo
39    <lexdo:editor rdf:resource="http://lexbib.org/agents/person/MeladzeGeorge",<
40    <dcterms:date>2016-01-01T02:00:00.000Z</dcterms:date>
41    <lexdo:publicationLanguage rdf:resource="http://lexvo.org/id/iso639-3/eng">
42    <lexdo:event rdf:resource="http://lexbib.org/events/ConfEuralex2016"/>
43    <lexdo:container rdf:resource="http://worldcat.org/isbn/9789941135422"/>
44    <lexdo:collection>2</lexdo:collection>
45    <lexdo:abstract rdf:resource="https://euralex.org/publications/on-bullying-mo
46    <lexdo:fullTextUrl>https://euralex.org/publications/on-bullying-mobbing-and-har
47    <lexdo:zoteroItemUri rdf:resource="http://zotero.org/groups/1892855/items/M9Z45UVF"
48    <lexdo:zoteroItemID>M9Z45UVF</lexdo:zoteroItemID>
49    <lexdo:firstAuLoc>http://en.wikipedia.org/wiki/Poznań</lexdo:firstAuLoc>
50    <rdf:type rdf:resource="http://lsdis.cs.uga.edu/projects/semdis/opus#Artic
```



Author name variants

```
lexperson:AtkinsBTS a lexdo:Person ;  
    skosxl:prefLabel [ a skosxl:Label ;  
        lexdo:nameVarFreq 3 ;  
        skosxl:literalForm "B. T. S. Atkins" ;  
        foaf:firstName "B. T. S." ;  
        foaf:surname "Atkins" ] .
```



skos:altLabel

```
lexperson:AtkinsBTsue a lexdo:Person ;  
    skosxl:prefLabel [ a skosxl:Label ;  
        lexdo:nameVarFreq 5 ;  
        skosxl:literalForm "B. T. Sue Atkins" ;  
        foaf:firstName "B. T. Sue" ;  
        foaf:surname "Atkins" ] .
```



skos:prefLabel

```
lexperson:AtkinsBerylT a lexdo:Person ;  
    skosxl:prefLabel [ a skosxl:Label ;  
        lexdo:nameVarFreq 1 ;  
        skosxl:literalForm "Beryl T. Atkins" ;  
        foaf:firstName "Beryl T." ;  
        foaf:surname "Atkins" ] .
```



skos:altLabel



Preparing authors for LOD-ification

- ▶ Harmonize name variants
 - ▶ Assign preferred name form
 - ▶ Detect valid name variants
 - ▶ Detect errors

- ▶ Disambiguate authors
 - ▶ Identify different persons with identical name literals
 - ▶ Namesakes, spelling errors, transliteration or encoding issues, ...

	rdfs_label	foaf_firs
1.	B. T. Sue Atkins	B. T. Sue
2.	Beryl T. Sue Atkins	Beryl T. Sue
3.	Sue Atkins	Sue
4.	Sue B. T. Atkins	Sue B. T.
5.	Miran Atkins	Miran



Assessing reference datasets

▼ Potential linking targets



▼ Assessment criteria

- ▼ Coverage
- ▼ Data quality
- ▼ Data structure
- ▼ Data modeling
- ▼ Interlinking within LOD Cloud

Harmonizing author names with OpenRefine

Clustering methods for duplicate detection

Show as: rows records Show: 5 10 25 50 rows			
All	rdfs_label	foaf_firstname	foaf_surname
1.	B. T. Sue Atkins	B. T. Sue	Atkins
2.	Beryl T. Sue Atkins	Beryl T. Sue	Atkins
3.	Sue Atkins	Sue	Atkins
4.	Sue B. T. Atkins	Sue B. T.	Atkins
5.	Miren Azkarate	Miren	Azkarate
6.	Laura Giacomini	Laura	Giacomini

Method key collision

Keying Function fingerprint

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	2	<ul style="list-style-type: none">B. T. Sue Atkins (1 rows)Sue B. T. Atkins (1 rows)	<input checked="" type="checkbox"/>	B. T. Sue Atkins

Clustering algorithms

- Rely entirely on string matching techniques (based on characters, n-grams, or phonetics)
- Settings range from conservative to liberal
- Detailed documentation:
<https://github.com/OpenRefine/OpenRefine/wiki/Clustering>



Harmonizing author names with OpenRefine

Creating records from clustering results

		Show as: rows records	Show: 5 10 25 50 records	
	All	rdfs_label	foaf_firstname	foaf_surname
	1.	B. T. Sue Atkins	B. T. Sue	Atkins
			Sue B. T.	Atkins
	2.	Beryl T. Sue Atkins	Beryl T. Sue	Atkins
	3.	Miren Azkarate	Miren	Azkarate
	4.	Sue Atkins	Sue	Atkins



Reconciling author names with Wikidata

sons_showcase csv [Permalink](#)

/ 7

Remove All

change

rs change reset

Blank Error 0

4 records

Show as: rows records Show: 5 10 25 50 records

	All	rdfs_label	foaf_firstname	foaf_surname
1.	B. T. Sue Atkins	<input checked="" type="checkbox"/> B. T. S. Atkins (92) <input checked="" type="checkbox"/> Create new item Search for match	edit	B. T. Sue Atkins
2.	B. T. S. Atkins	Choose new match		
3.	Miren Azkarate Villar	Choose new match		
4.	Sue Atkins	<input checked="" type="checkbox"/> B. T. S. Atkins (100) <input checked="" type="checkbox"/> Sue Atkins (100) <input checked="" type="checkbox"/> Sue Atkinson (91) <input checked="" type="checkbox"/> Create new item Search for match		

Match this Cell Match All Identical Cells Cancel

B. T. S. Atkins (Q4834227)
British lexicographer and linguist (*1930) ♀



Reconciliation services for OpenRefine: <https://reconciliation-api.github.io/testbench/>



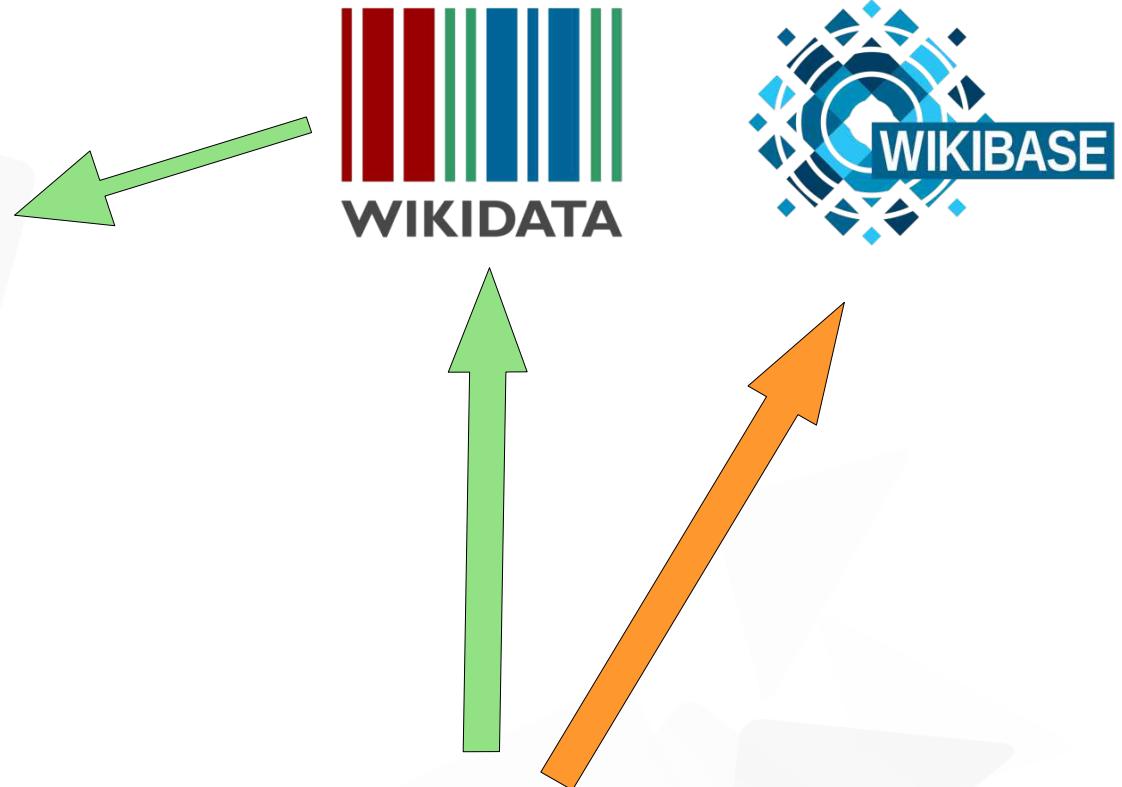
Exporting from OpenRefine

Sue B. T. Atkins
B. T. S. Atkins (92)
Create new item

Search for match

Miren Azkarate Villar
Choose new match

Laura Giacomini
Create new item



Sue Atkinson (91)
Create new item

Sue Atkinson (91)
Create new item

Sue Atkinson (91)
Create new item



Linking results for LexBib persons with OpenRefine

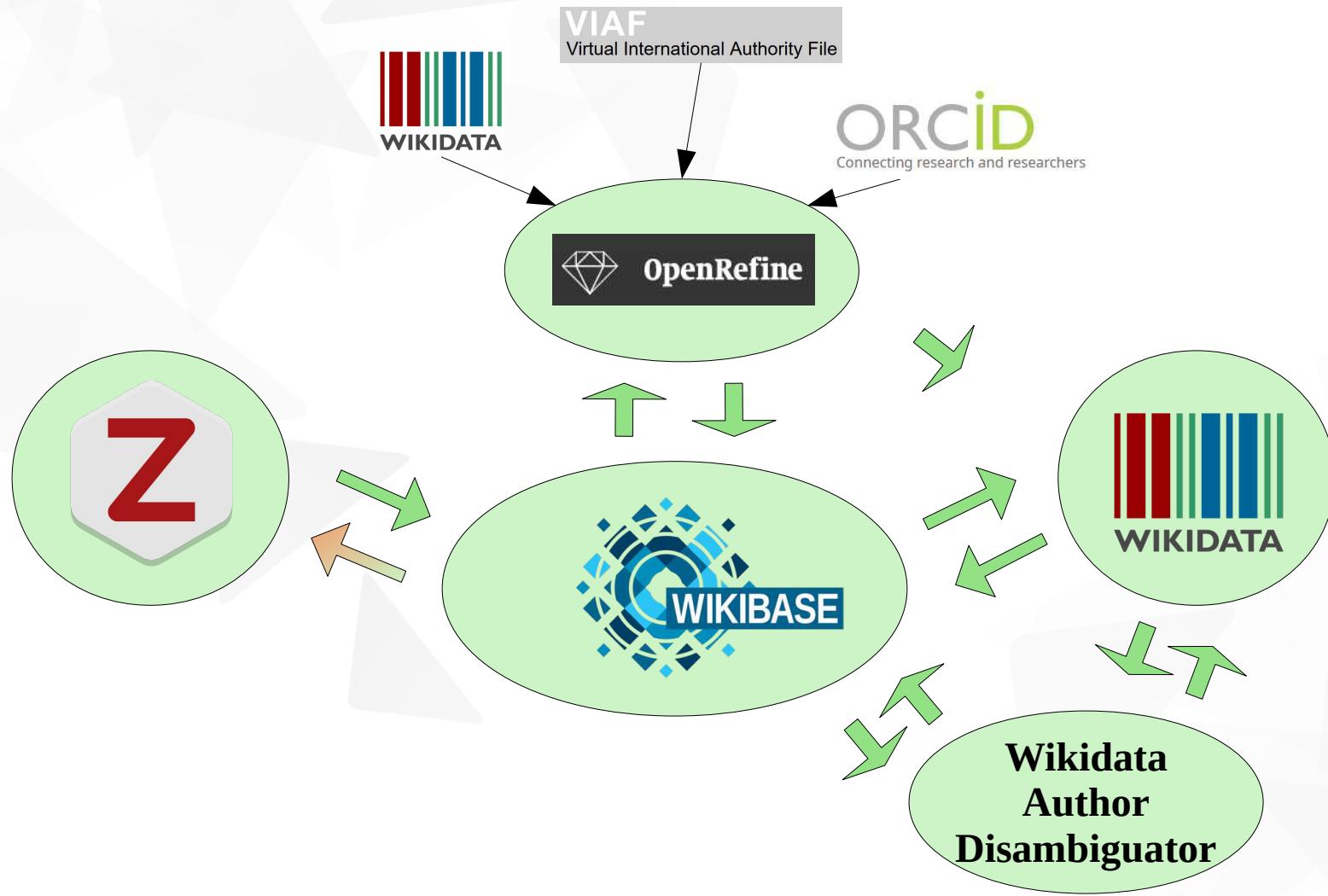
- ▼ Sample: 2.923 authors
- ▼ Validated: 100 automatic matches + 100 matching candidates

	 WIKIDATA	VIAF Virtual International Authority File		
	% (complete sample)	Precision (validated subset)	% (complete sample)	Precision (validated subset)
Automatic match	22	0.9	40	0.86
Candidates (score: 0.8 – 1.0)	14	0.42	17	0.95
No match found	64	Not assessed	43	Not assessed



- straightforward
- not trivial but possible
- not possible

Data migration scheme using Wikibase



Wikidata, and own Wikibase

Wikidata

- ▶ Publication metadata
- ▶ Places
- ▶ Organizations
- ▶ Persons
- ▶ Events
- ▶ Projects
- ▶ Languages



Own wikibase instance

- ▶ as sandbox
- ▶ as collaborative workbench
- ▶ for instant data exhibits and visualisations
- ▶ for harvesting from and feeding to wikidata
- ▶ May replace actual RDF database



Related work:
OCLC Project "Passage"



Item Discussion

Euralex 1990 (Q100594659)

Academic conference about Lexicography

▼ In more languages

Language	Label	Description
English	Euralex 1990	Acader
German	No label defined	No des
Spanish	No label defined	No des
Basque	No label defined	No des
Latin	No label defined	No des

Statements

instance of academic conference

▼ 0 references

part of Euralex conference series

▼ 0 references

location Málaga

▼ 0 references



Tools and Related Work

▼ Discussed tools

- ▶ Zotero
- ▶ VocBench3
- ▶ Ontotext RDF Triple Store (not open source)
- ▶ Wikibase
- ▶ ZotKat and Quickstatements
- ▶ OpenRefine
- ▶ Wikidata Author Disambiguator
- ▶ Zotero WikiCite Plugin (planned)

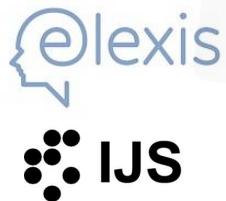
▼ Related work

- ▶ Bibliodata as LOD
 - ▶ OCLC Project "Passage"
 - ▶ Linked.swissbib
 - ▶ IFLA Best Practices for Bibliodata as LOD
- ▶ Single-domain databases, incl. bibliodata
 - ▶ FactGrid Database
- ▶ Wikidata-based scholarly profiles
 - ▶ Scholia



Outlook

- ▼ Ongoing work on LexBib collection and LexDo domain ontology
- ▼ DARIAH WG Bibliodata activities
 - ▼ Feedback
 - ▼ Collaboration
 - ▼ Development of tool surveys and best practice guidelines
 - ▼ Structuring of plain text bibliographies
 - ▼ Bibliographical data as LOD
 - ▼ Ontological representation of scientific domains
 - ▼ Free open source tools, semantic web standards, and developing communities



This project has received funding from
IT1169-19 Consolidated Research Group of Excellence
(2019-2021), Basque Government



Contact: david@lexbib.org / christiane@lexbib.org

ResearchGate: <https://www.researchgate.net/project/LexBib-Corpus-and-Bibliography>



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

DARIAH Annual Event 2020: WG Bibliodata Workshop