



Business and Economics Research Data Center

BERD@BW

Baden-Württemberg

<https://www.berd-bw.de>

Wikibase knowledge graphs for data management & data science

Dr. Renat Shigapov

23.06.2021



@shigapov



@_shigapov

Motivation

**DATA
MANAGEMENT**

1. people

**DATA
SCIENCE**

2. processes

3. technology

**KNOWLEDGE
GRAPHS**



knowledge
information
data

**linking
things**



Flow

- Definitions
- Wikidata & Tools
- Local Wikibase
- Wikibase Ecosystem
- Summary

**DATA
SCIENCE**

**DATA
MANAGEMENT**



**KNOWLEDGE
GRAPHS**

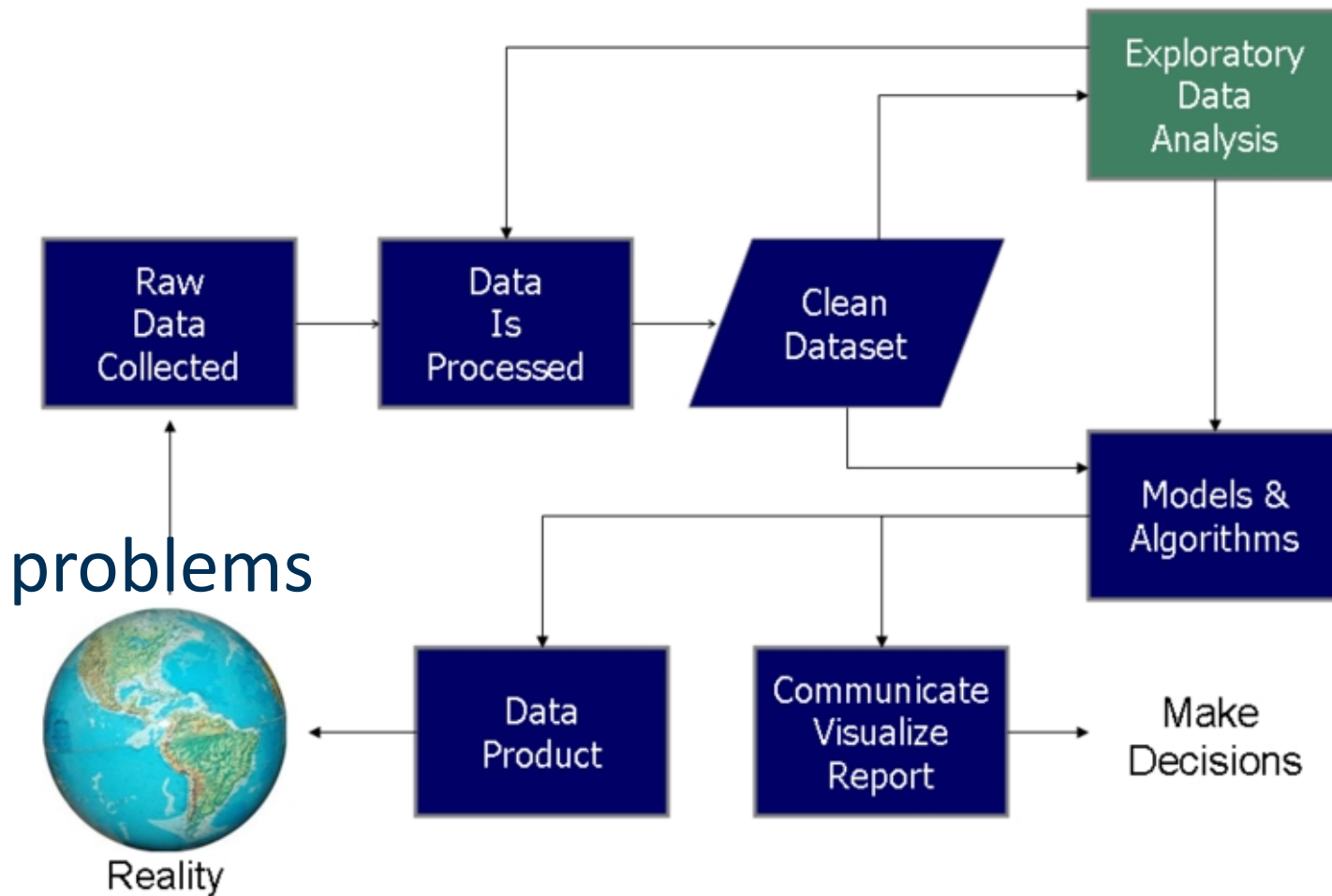
29.10.2012

2021

2030

DATA SCIENCE

Data Science Process



Example: Named Entity Linking

“Paris is the capital of France”

wikipedia.org/wiki/Paris

wikipedia.org/wiki/France

https://commons.wikimedia.org/wiki/File:Entity_Linking_-_Short_Example.png

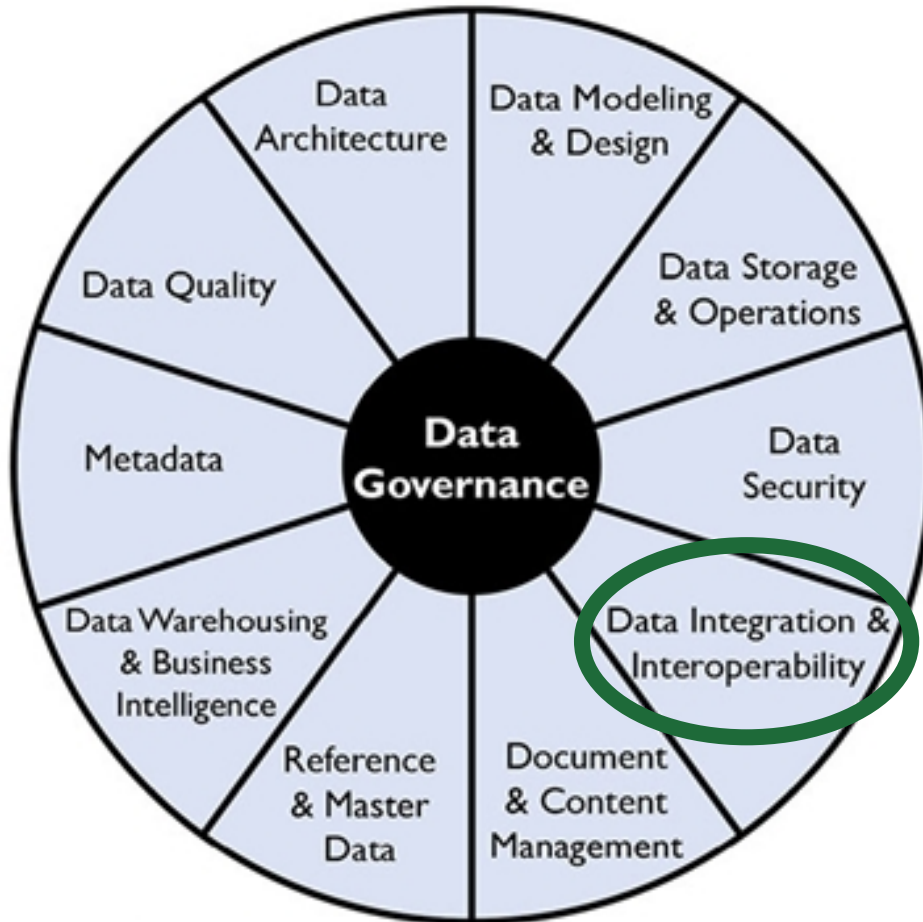
Rule-based
Machine Learning
Deep Learning

Learn data science at

<https://www.kaggle.com>

DATA MANAGEMENT

general



DAMA-DMBOK2 Data Management Framework

Copyright © 2017 by DAMA International

data silos

data fabric

data mesh

data space

data marketplace

data lake

data swamp

Example:

research



Research data lifecycle



+



**KNOWLEDGE
GRAPH**

=

**ONTOLOGY
+ THINGS**

<https://www.mediawiki.org>

<https://www.wikiba.se>



+



<https://mariadb.org>

<https://blazegraph.com>

relational database

graph database

- ✓ “Things, not strings” by Google, 2012
- ✓ A knowledge graph **links** things in different datasets
- ✓ A knowledge graph **can link** people & processes and **enhance** technologies

The main
example:



<https://www.wikidata.org>

“THE KNOWLEDGE GRAPH COOKBOOK RECIPES THAT WORK” by
ANDREAS BLUMAUER & HELMUT NAGY, 2020.

Wikidata in 2012: The start of big data integration

motivation



To link unlinked,
unstructured, multilingual &
very dynamic data

agile collaborative solution



<https://www.wikidata.org>

Links structured & unstructured data
Can be edited by humans & machines

29.10.2012

Watch talks by Lydia Pintscher at SMWCon Falls 2013, 2016 & 2020:
https://www.semantic-mediawiki.org/wiki/User:Lydia_Pintscher

Wikidata in 2021: Data management with a knowledge graph works

links 94+ millions entities (things)
in 6200+ datasets around the world

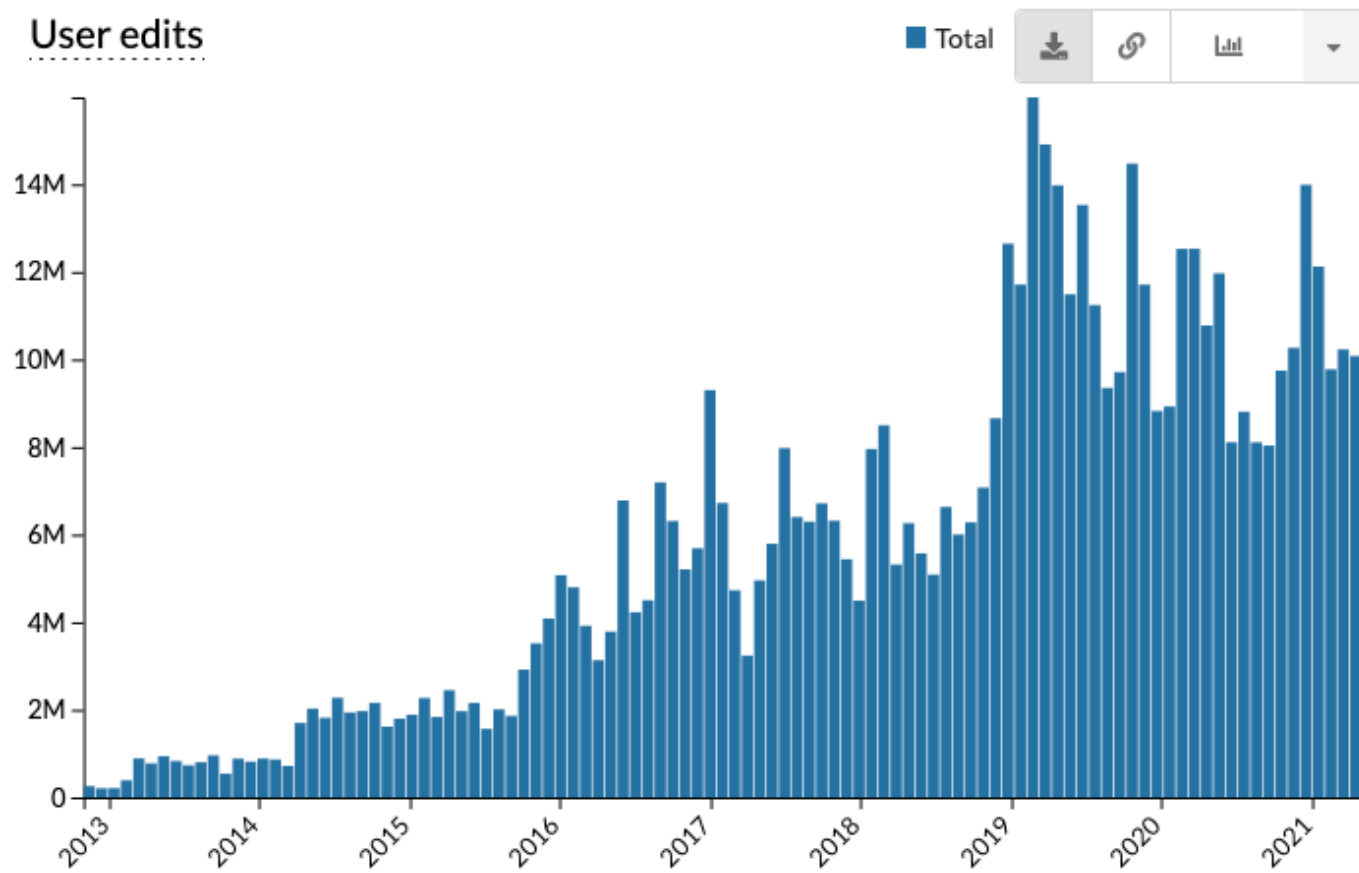


23.06.2021

How did it work out?

people via the Wikidata frontend

User edits



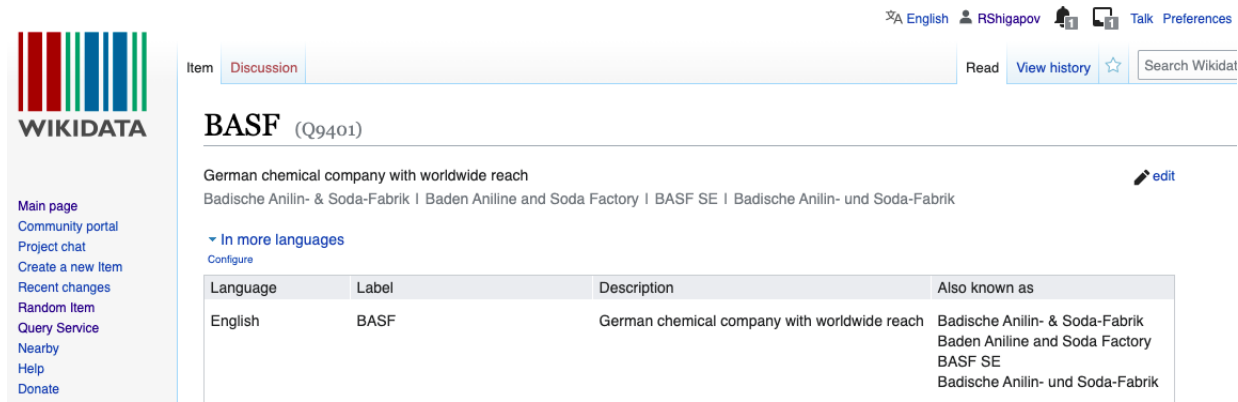
bots via the Wikidata API

Top editors

Edits	Name
1,373,426	Edoderoobot
1,336,162	Lockalbot
1,195,664	YoaR
1,160,771	Citationgraph bot
1,109,029	SuccuBot

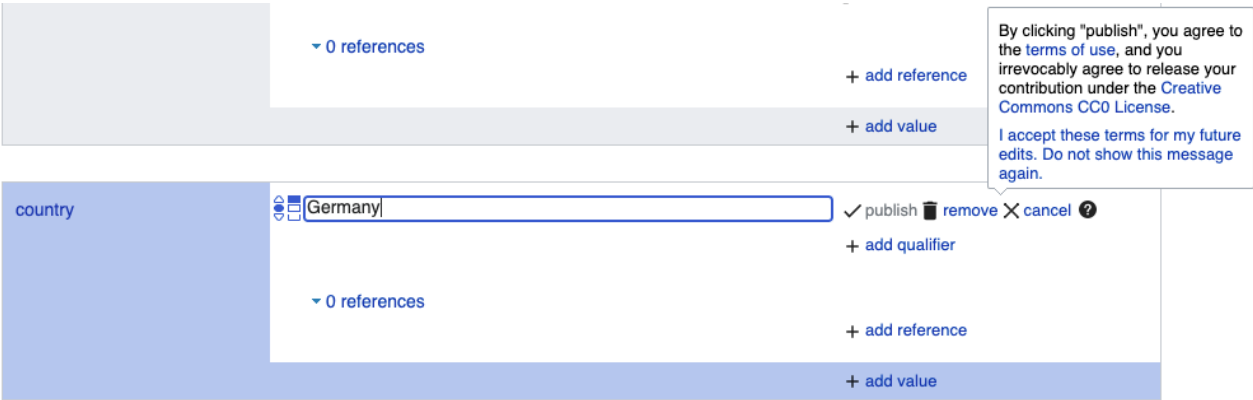
Tools for data import

the Wikidata frontend



the Wikidata API and its wrappers

many options



Tools for data import: the wrappers of the Wikidata API

Maxime Lathuilière
(aka maxlath):

<https://github.com/maxlath/wikibase-edit>

<https://github.com/maxlath/wikibase-cli>

Markus Krötzsch and Co:

<https://github.com/Wikidata/Wikidata-Toolkit>

Wikimedia:

<https://github.com/wikimedia/pywikibot>

OpenRefine

GUI

Magnus Manske:

GUI

<https://github.com/magnusmanske/quickstatements>

Andra Waagmeester and Co:

<https://github.com/SuLab/WikidataIntegrator>

LeMyst:

<https://github.com/LeMyst/WikibaseIntegrator>

Tabular data cleaning & Wikidata reconciliation services

OpenRefine



3.	16765	Universidad de Zaragoza	☒ ☒ Universidad de Zaragoza (100) ☒ Create new topic Search for match
4.	16778	Universidad de Sevilla	☒ ☒ Universidad de Sevilla (100) ☒ Create new topic Search for match
5.	2152	University of Cambridge	☒ ☒ University of Cambridge ☒ University of Cambridge ☒ Create new topic Search for match
6.	1259	University of Michigan	Choose new match
7.	14736	Texas A&M University	Choose new match
8.	1877	University of Colorado Boulder	☒ ☒ University of Colorado Boulder (100) ☒ Create new topic Search for match
9.	72971	Universidade Técnica de Lisboa Instituto Superior Técnico	☒ ☒ Instituto Superior Técnico (63) ☒ Create new topic Search for match



<https://commons.wikimedia.org/w/index.php?curid=60388061>

<https://github.com/OpenRefine/OpenRefine>

Reconciliation service API

Browser screenshot showing the Reconciliation service API interface. The URL is <https://wikidata.reconci.link/en/api?query=BASF>. The interface displays JSON results for the query "BASF".

```
JSON Raw Data Headers
Save Copy Collapse All Expand All Filter JSON

▼ result:
  ▼ 0:
    ▼ features:
      ▼ 0:
        id: "all_labels"
        value: 100
      id: "Q9401"
      match: false
      name: "BASF"
      score: 100
    ▼ type:
      ▼ 0:
        id: "Q4830453"
        name: "business"
      ▼ 1:
        id: "Q6881511"
        name: "enterprise"
      ▼ 2:
        id: "Q891723"
        name: "public company"
```

<https://github.com/wetneb/openrefine-wikibase>

<https://github.com/reconciliation-api>

Named entity linking on tables & automatic ontology learning

Data science competition using the Wikidata knowledge graph:

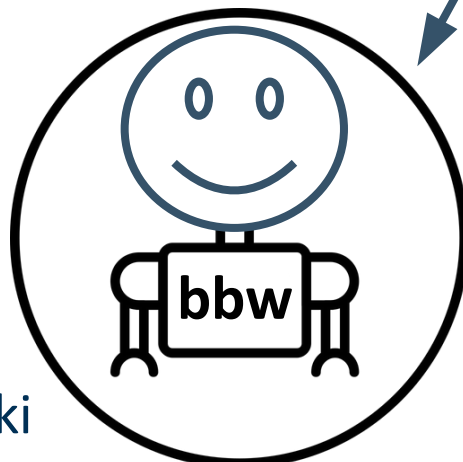
<http://www.cs.ox.ac.uk/isg/challenges/sem-tab/2020>

SemTab: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching

3rd Prize

bbw Team

Renat Shigapov, Philipp Zumstein, Jan Kamlah, Lars Oberländer, Jörg Mechnich, and Irene Schumm



Boosted By Wiki

1	Deutsche Telekom	mobile phone industry	Bonn	1994-01-01	https://www.telekom.com/
2	Lufthansa	air transport	Cologne	1953-01-01	https://www.lufthansa.com
3	SAP SE	software development	Walldorf	1972-04-01	https://www.sap.com
1	Deutsche Telekom	mobile phone industry	Bonn	1994-01-01	https://www.telekom.com/
2	Lufthansa	air transport	Cologne	1953-01-01	https://www.lufthansa.com
3	SAP SE	software development	Walldorf	1972-04-01	https://www.sap.com
property		industry	headquarters location	inception	official website
type	business	industry	urban municipality of Germany		
datatype		WikibaseItem	WikibaseItem	Time	Url

<https://github.com/UB-Mannheim/bbw>

Named entity linking on texts with Wikidata

Real-time NEL on Wikidata

[[OpenTapioca^β]]

OpenTapioca annotates text with locations, organizations and people from Wikidata.
It is synchronous with Wikidata.
This is a prototype. See the [GitHub project](#) for more info.

Paris is the capital of France.

Annotate

Paris is the capital of France.

Paris (Q90)
capital and largest city of France
Rank: 13.82, phrase: 10.31
Statements: 399, sitelinks: 326
Score: 2.5008074933329114

<https://github.com/wetneb/opentapioca>

SOTA: training on Wikipedia +
linking to Wikidata for free


BLINK

 GENRE

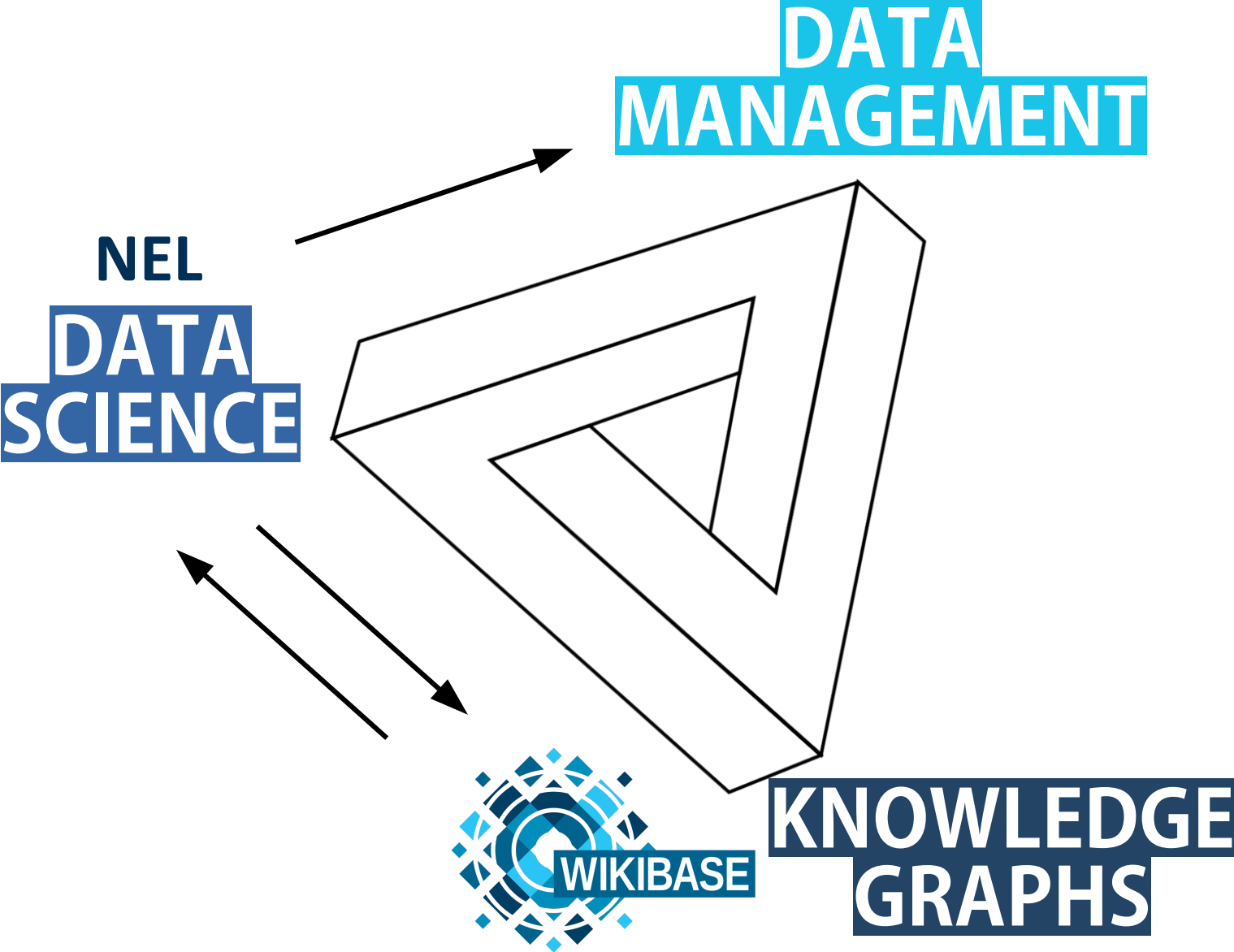
<https://github.com/facebookresearch/BLINK>

<https://github.com/facebookresearch/GENRE>

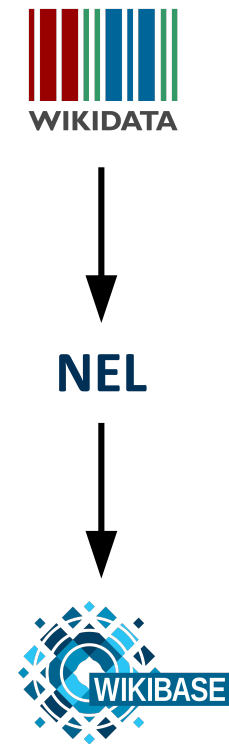
using database: not SOTA, but simple:

<https://github.com/egerber/spaCy-entity-linker>

Importance of Named Entity Linking (NEL)



Creating your own Wikibase knowledge graph:



The main Wikibase Knowledge Graph



an agile collaborative data integration
process connecting people & advancing
technology (e.g., data science)

Motivation for a local Wikibase Knowledge Graph



multiple unlinked
projects & datasets
containing info about
the same things



Semantic interoperability in

1. (research) data management
2. content management
3. knowledge management

A Wikibase Knowledge Graph from scratch: Installation

1. manual

**2. Docker
image**

docker-compose up

**3. WbStack
(Wikibase as a
service)**

<https://www.mediawiki.org/wiki/Wikibase/Docker>

<https://www.mediawiki.org/wiki/Wikibase/Installation>

<https://www.wbstack.com>

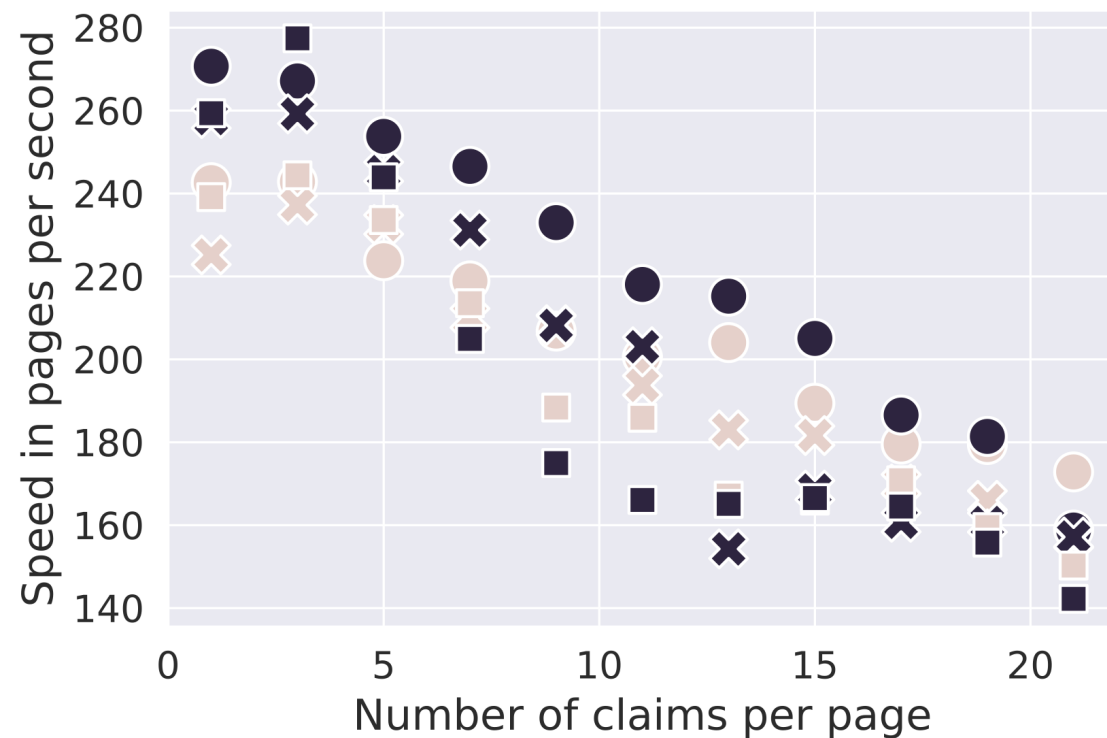
simplicity of installation

Data import into a local Wikibase instance: One more option via

Wikibase MariaDB



speed up!



<https://github.com/UB-Mannheim/RaiseWikibase>

BASF (Q9401)

[Configure](#)

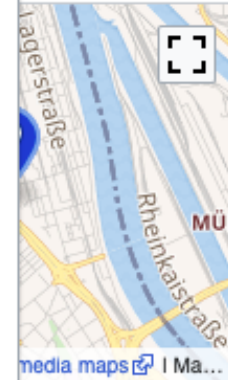
Language	Label
English	BASF



×

[Help](#) [Discuss](#)

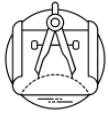
- postal code
- located on street
- street number
- located in the administrative territorial entity
- country
- location
- coordinate location
- start time
- end time
- street address
- ...



[Panorama.jpg](#)

Carl-Bosch-Strasse 38 (German)

Data Validation: EntitySchemas & ShEx validation



Wikidata Schemas

author (E42)

language code	label	description	aliases	edit
en	author	A Schema for authors		edit
ast	autor	esquema p'autores	autora	edit

```
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX ps: <http://www.wikidata.org/prop/statement/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX psn: <http://www.wikidata.org/prop/statement/value-normalized/>

# Example queries:
# 5 authors: SELECT * WHERE { ?person wdt:P106 wd:Q36180 } LIMIT 5
# Specific author (Q42): SELECT DISTINCT ?author WHERE { VALUES ?author { wd:Q42 } }

start = @<AuthorShape>

<AuthorShape> EXTRA wdt:P31 wdt:P106 {

  # instance of
  wdt:P31 [ wd:Q5 ] ;

  # given name
  wdt:P735 @<GivenName> * ;

  # sex or gender
  wdt:P21 IRI ?;

  # family name
  # wdt:P734 IRI ?;

  # occupation
  wdt:P106 [ wd:Q36180 ] ;

  # Commons category
  wdt:P373 . ?;

  # notable work
  wdt:P800 IRI *;
```

ShEx2 — Simple Online Validator

```
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX ps: <http://www.wikidata.org/prop/statement/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX psn: <http://www.wikidata.org/prop/statement/value-normalized/>

# Example queries:
# 5 authors: SELECT * WHERE { ?person wdt:P106 wd:Q36180 } LIMIT 5
# Specific author (Q42): SELECT DISTINCT ?author WHERE { VALUES ?author { wd:Q42 } }

start = @<AuthorShape>

<AuthorShape> EXTRA wdt:P31 wdt:P106 {

  # instance of
  wdt:P31 [ wd:Q5 ] ;

  # given name
  wdt:P735 @<GivenName> * ;

  # sex or gender
  wdt:P21 IRI ?;
```

The lecture by Jose Emilio Labra Gayo and Andra Waagmeester in the Stanford Course on Knowledge Graphs:

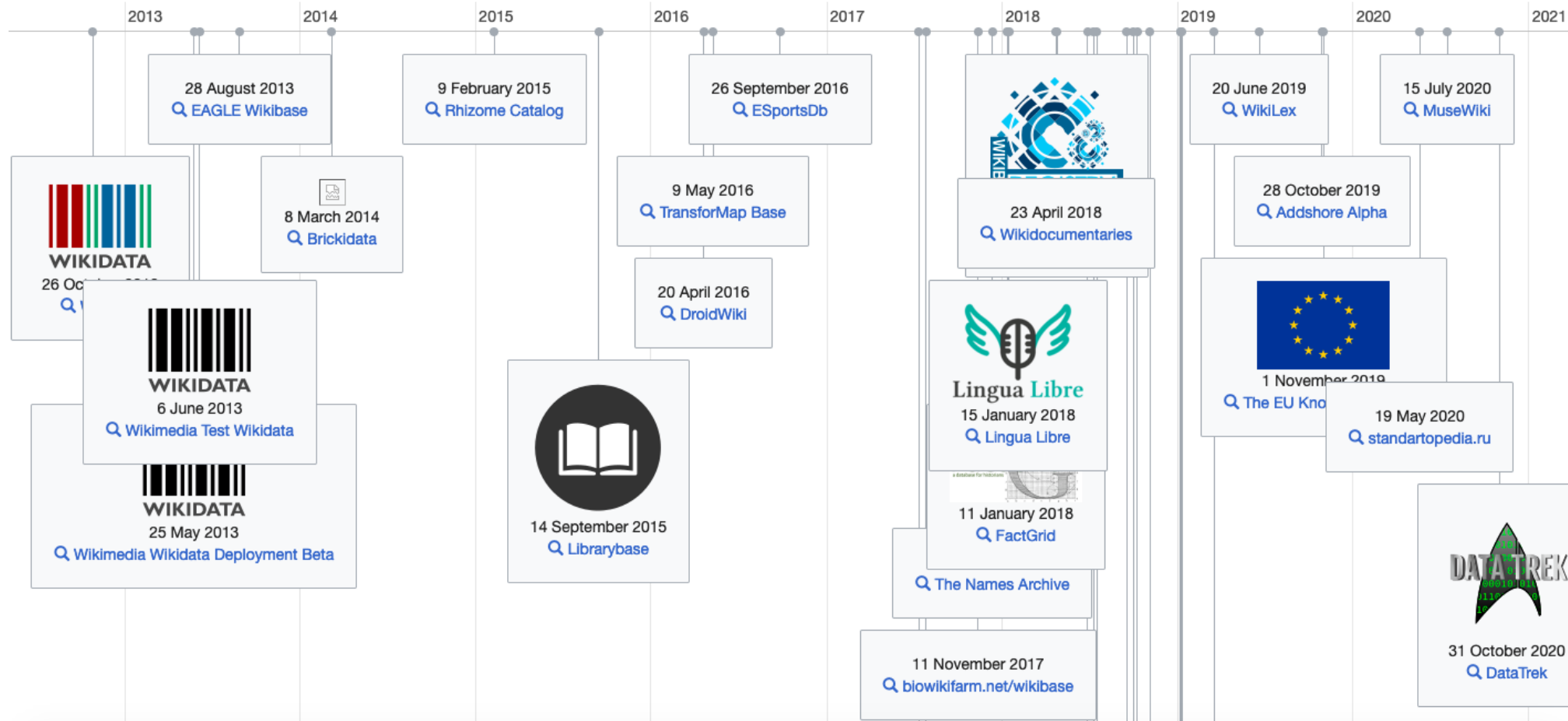
<https://youtu.be/IE1ZF02-yIO?t=1860>

<https://web.stanford.edu/class/cs520/abstracts/gayo-waagmeester>

<https://www.wikidata.org/wiki/EntitySchema:E42>

See references for “A protocol for adding knowledge to Wikidata” 21

Towards the Wikibase Ecosystem

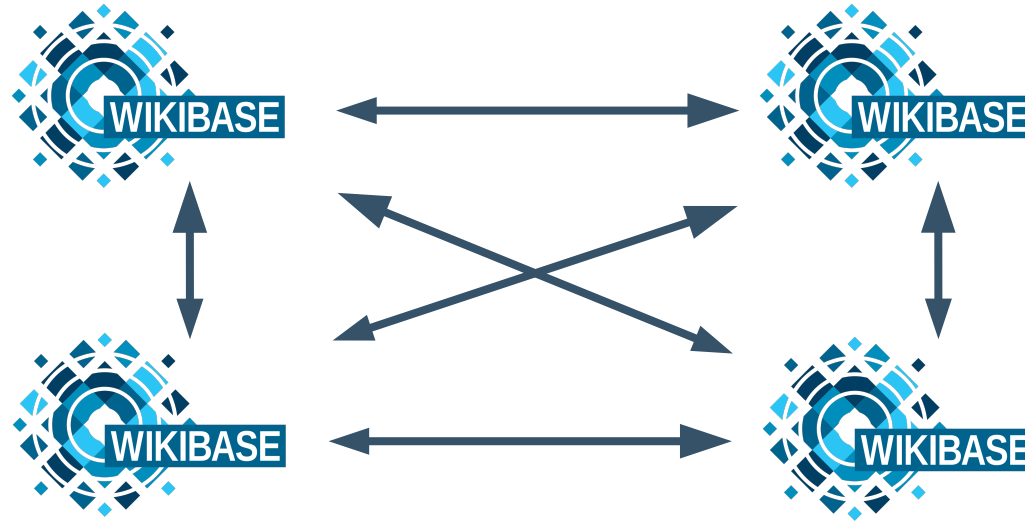


The Wikibase Registry has been launched by Adam Shorland: <https://wikibase-registry.wmflabs.org>

“The strategy for Wikibase Ecosystem” by Lydia Pintscher et al. at
https://upload.wikimedia.org/wikipedia/commons/c/cc/Strategy_for_Wikibase_Ecosystem.pdf

Challenges

Ontology reuse (not only **federated properties** & **constraints** from Wikidata, but **any** ontology)



Even more knowledge sharing among the Wikibase maintainers (tutorials, use cases, papers & codes)

Reuse of **any** Wikidata-specific software (WikibaseManifest?): data import & validation, data scientific & monitoring tools

Towards all-in-one data management solution maintained by one person

Community of practice

Wikidata & Wikibase Office hours

WikidataCon

Wikidata Workshop at ISWC

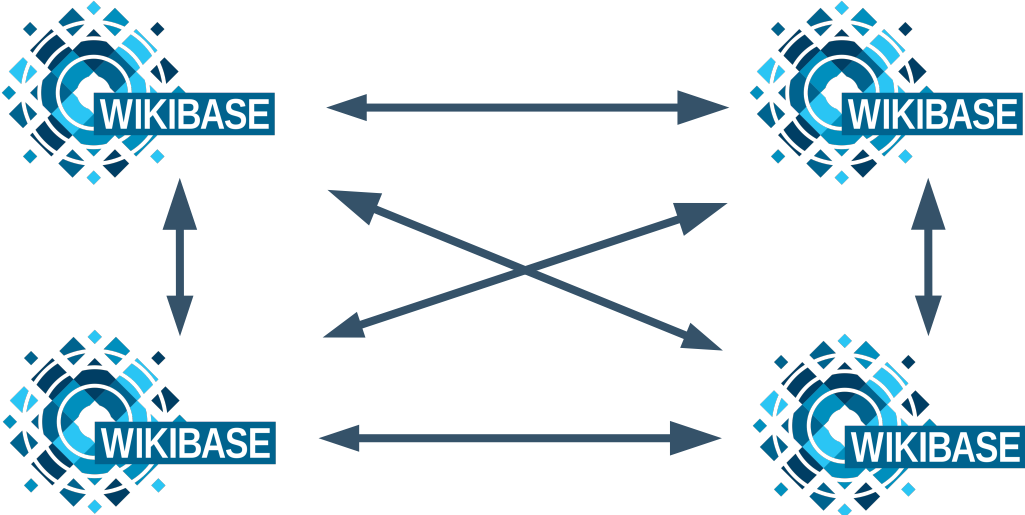
Wikidata telegram group

ISWC & ESWC

Wikibase Community telegram group

Wikibase Stakeholder Group

Wikidata bug triage hour



Summary

collaborative

**DATA
MANAGEMENT**

connect people

boost **DATA
SCIENCE**

link processes

enhance technology



**KNOWLEDGE
GRAPHS**

**link
things**

References

1. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledgebase. Commun. ACM 57(10), 7885 (Sep 2014), <https://doi.org/10.1145/2629489>
2. Delpuch, A. OpenTapioca: Lightweight Entity Linking for Wikidata, in Proceeding of Wikidata Workshop 2020, <http://ceur-ws.org/Vol-2773/paper-02.pdf>
3. Delpuch, A., Running a reconciliation service for Wikidata, in Proceeding of Wikidata Workshop 2020, <http://ceur-ws.org/Vol-2773/paper-17.pdf>
4. Waagmeester, A., Willighagen, E.L., Su, A.I. et al. A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses. BMC Biol 19, 12 (2021). <https://doi.org/10.1186/s12915-020-00940-y>
5. Burgstaller-Muehlbacher, S., et al. SuLab/WikidataIntegrator 0.5.1 (2020), <https://doi.org/10.5281/zenodo.3621065>
6. Waagmeester, A., et al. A protocol for adding knowledge to Wikidata, a case report, in bioRxiv, <https://doi.org/10.1101/2020.04.05.026336>
7. Pintscher, L., Voget, L., Koeppen, M., Aleynikova, E.: Strategy for the Wikibase Ecosystem (2019), <https://w.wiki/334L>
8. Shigapov, R., Zumstein, P., Kamlah, J., Oberländer, L., Mechnich, J., & Schumm, I. (2020). bbw: Matching CSV to Wikidata via Meta-lookup. In CEUR Workshop Proceedings, <http://ceur-ws.org/Vol-2775/paper2.pdf>
9. Shigapov, R., Mechnich, J. & Schumm, I. RaiseWikibase: Fast inserts into the BERD instance. ESWC 2021 Satellite Events, 2021, <https://openreview.net/pdf?id=87hp7LJDJE>