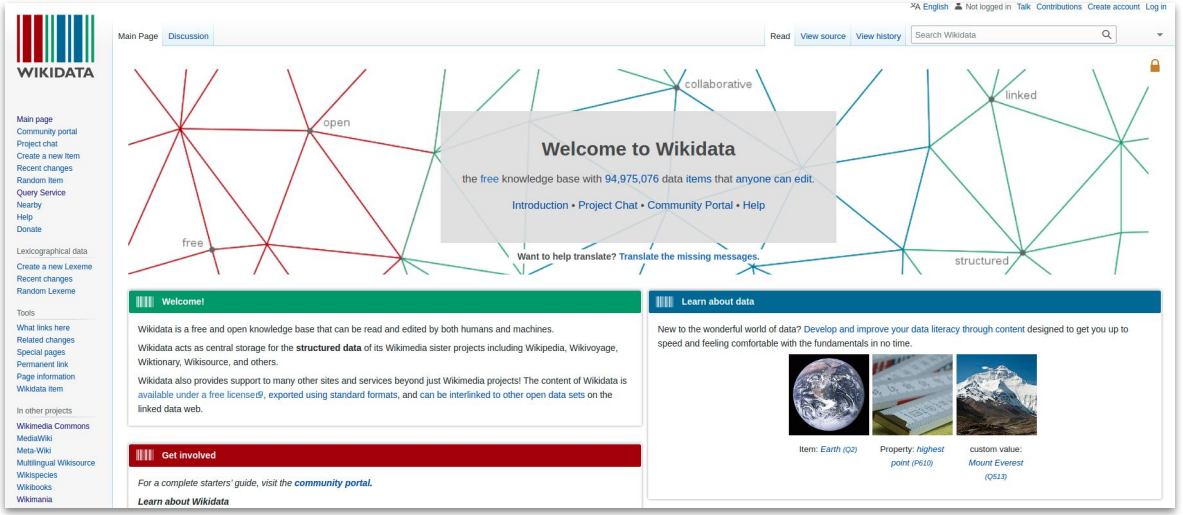
Lexicographical data in Wikidata: Because technology should speak your language!

Lydia Pintscher
Wikidata Portfolio Lead, Wikimedia Deutschland
lydia.pintscher@wikimedia.de - @nightrose
First Wikibase Lexical Data Workshop

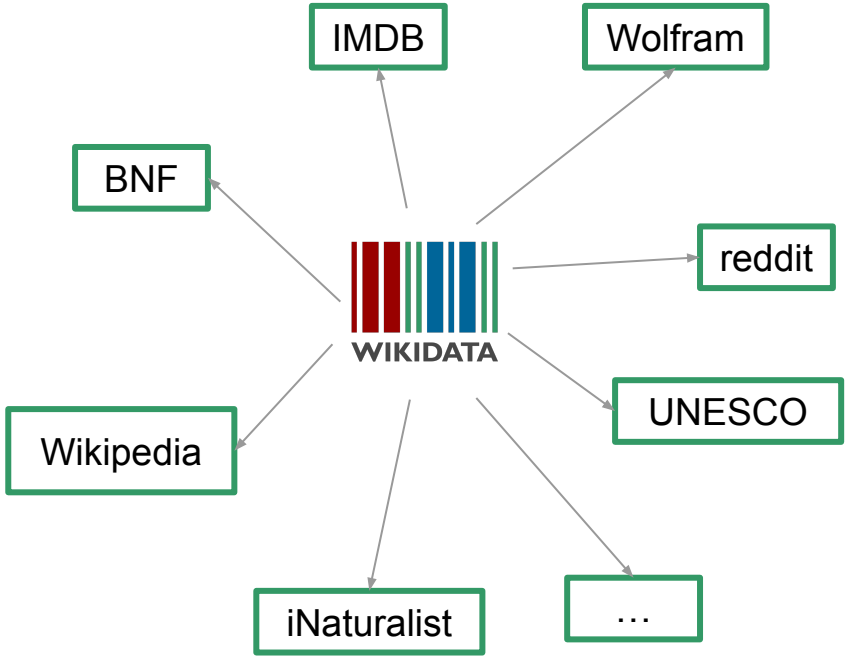
Wikidata today

- One of the largest Wikimedia projects
- Free and open knowledge graph
- Data is used in a lot of technology you use every day
- Data available under CC0
- Made for humans and machines
- Multilingual
- Collaborative



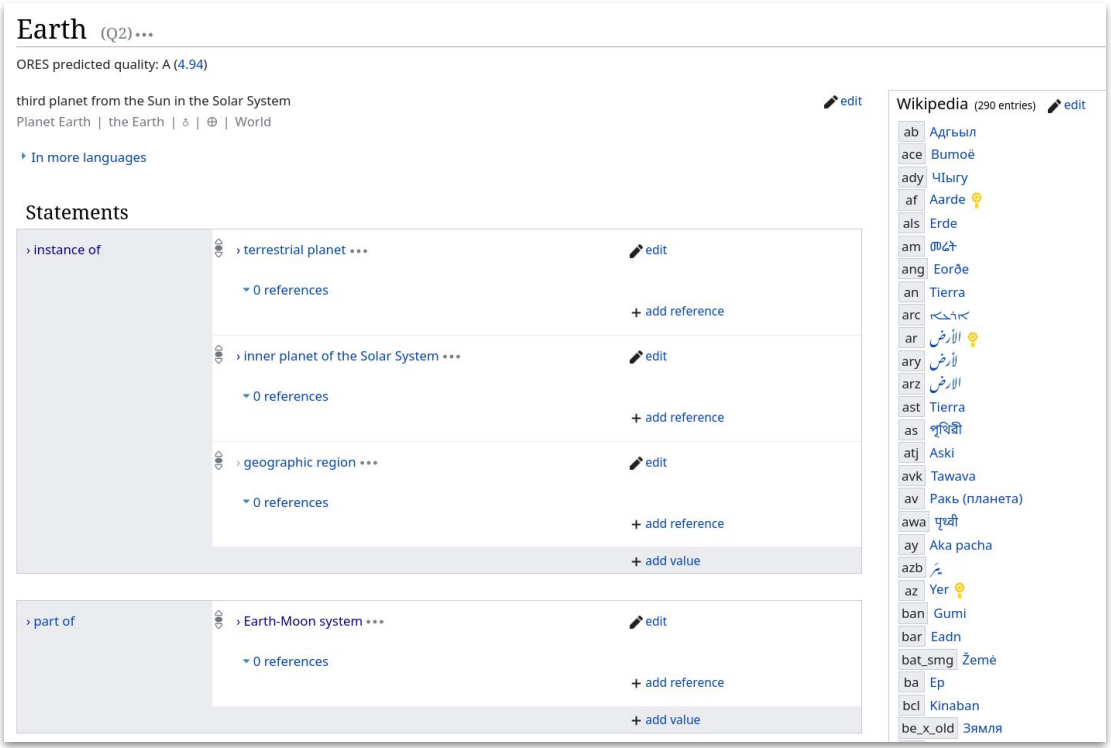
What makes Wikidata special?

- You can be a part of it
- More nuanced modeling of the world and focusing on verifiability
- Multilingual
- Loosely enforced ontology
- Highly connected internally and to other databases, catalogs, etc. to open up a ton of additional data
- Closely connected to Wikipedia and the other Wikimedia Projects



106 Million

Items



Earth (Q2)

ORES predicted quality: A (4.94)

third planet from the Sun in the Solar System

Planet Earth | the Earth | ♁ | ⊕ | World

In more languages

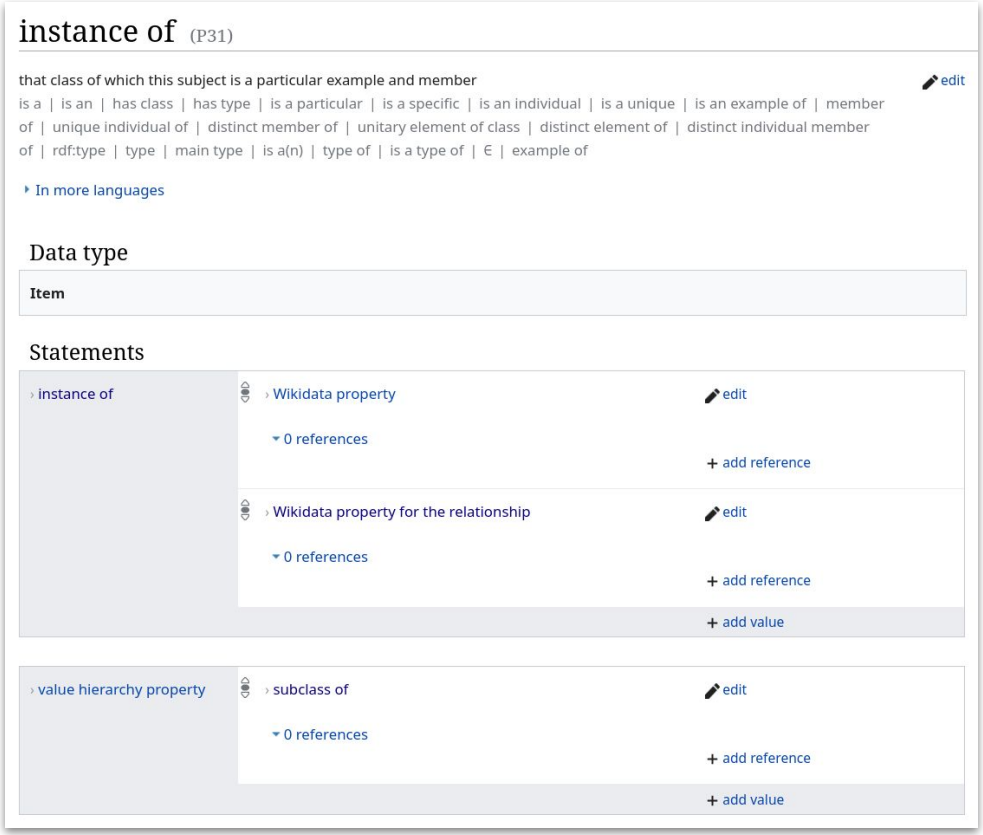
Statements

› instance of
› terrestrial planet
0 references
edit
add reference
› inner planet of the Solar System
0 references
edit
add reference
› geographic region
0 references
edit
add reference
add value

› part of
› Earth-Moon system
0 references
edit
add reference
add value

Wikipedia (290 entries) edit

ab Адгьыл
ace Bumoë
ady ЧIыгу
af Aarde
als Erde
am ሙሪት
ang Eorðe
an Tierra
arc ארעא
ar الأرض
ary الأرض
arz الارض
ast Tierra
as পৃথিৱী
atj Aski
avk Tawava
av Ракь (планета)
awa पृथ्वी
ay Aka pacha
azb یئر
az Yer
ban Gumi
bar Eadn
bat_smg Žemė
ba Ер
bcl Kinaban
be_x_old Зямля

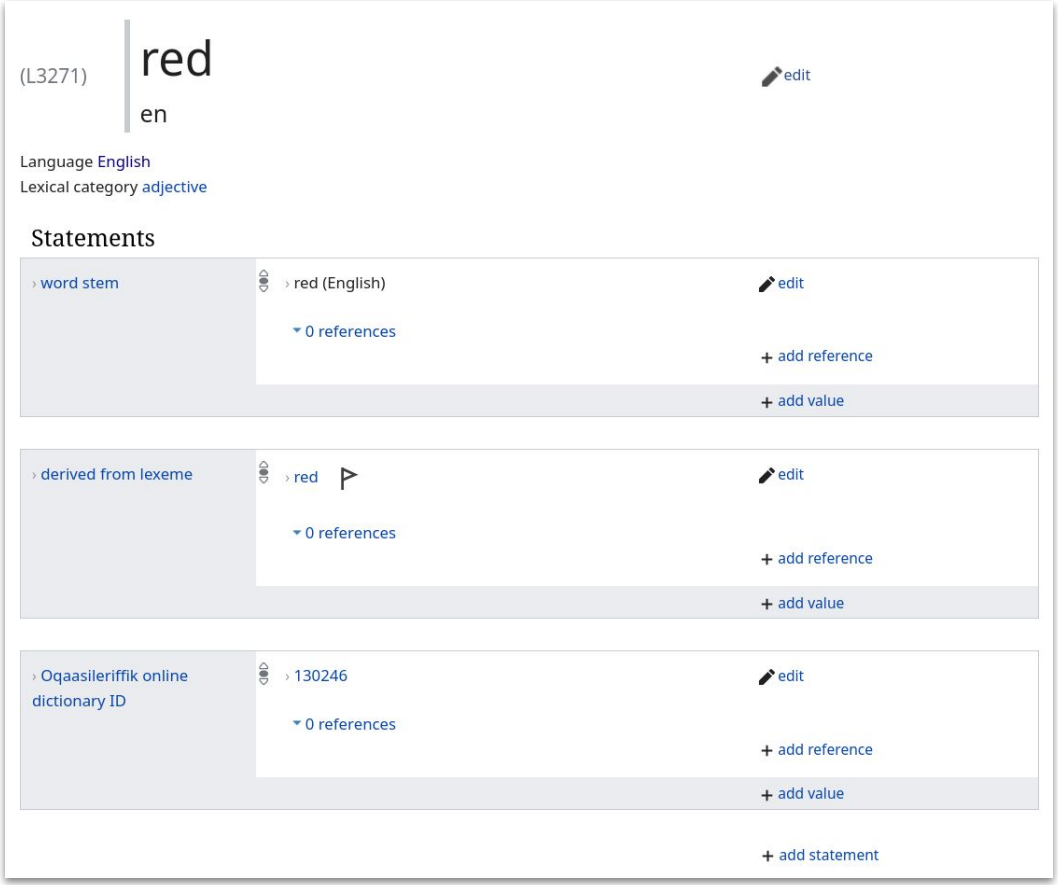


11k

Properties

1.51 Billion

Statements



1.14 Million

Lexemes

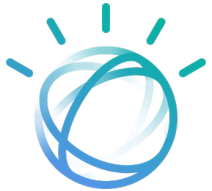
12.5k

active editors

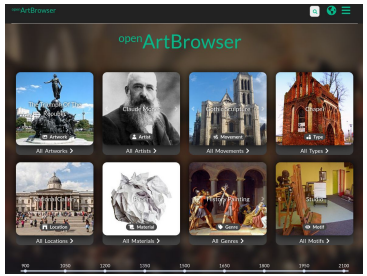
Photo by Pierre-Selim Huard (CC BY 4.0)

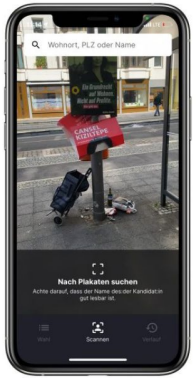

OPEN
LIBRARY


mySociety

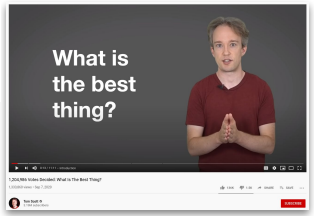



Google


openArtBrowser

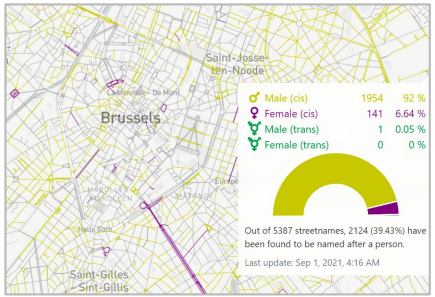

Wohnort, PLZ oder Name
Nach Plakaten suchen
Scannen


OCCRP


What is
the best
thing?

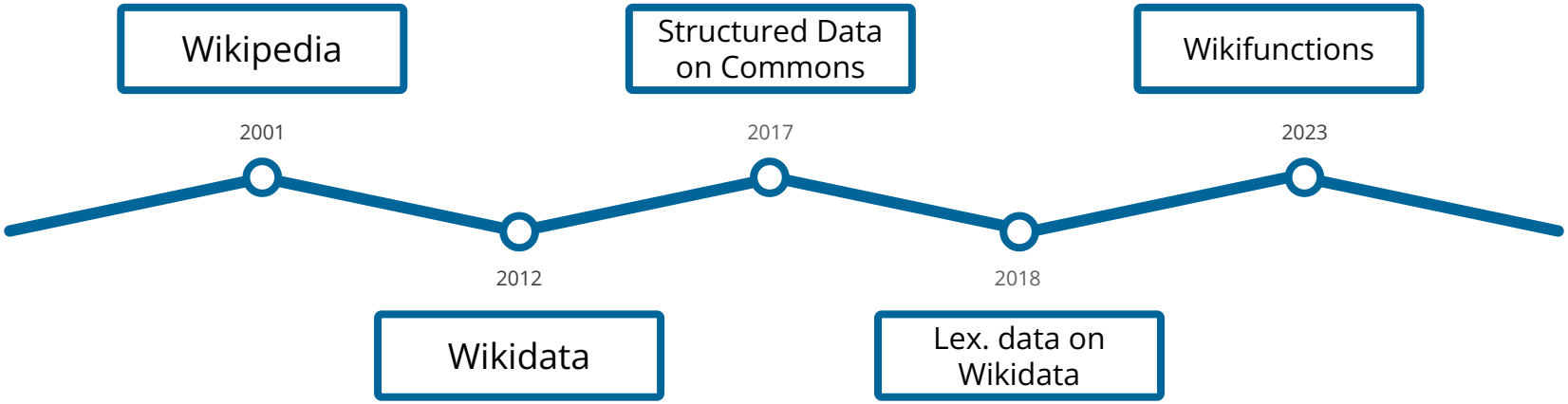

Quora


reddit


Saint-Josse-ten-Noode
Brussels
Saint-Gilles
Sint-Gillis
Male (cis) 1954 92 %
Female (cis) 141 6.64 %
Male (trans) 1 0.05 %
Female (trans) 0 0 %
Out of 5387 streetnames, 2124 (39.43%) have been found to be named after a person.
Last update: Sep 1, 2021, 4:16 AM


WolframAlpha™

How did we get here?


Wikipedia
2001
Wikidata
2012
Structured Data on Commons
2017
Lex. data on Wikidata
2018
Wikifunctions
2023

Envisioned use cases back then

- Support for other Wikimedia projects (e.g. Wiktionary)
- Dictionary apps
- Language learning tools (e.g. word lists)
- Research
- Text analysis (e.g. sentiment analysis)
- Text generation (e.g. for Abstract Wikipedia)

… Because technology should speak your language!

Why does this matter?

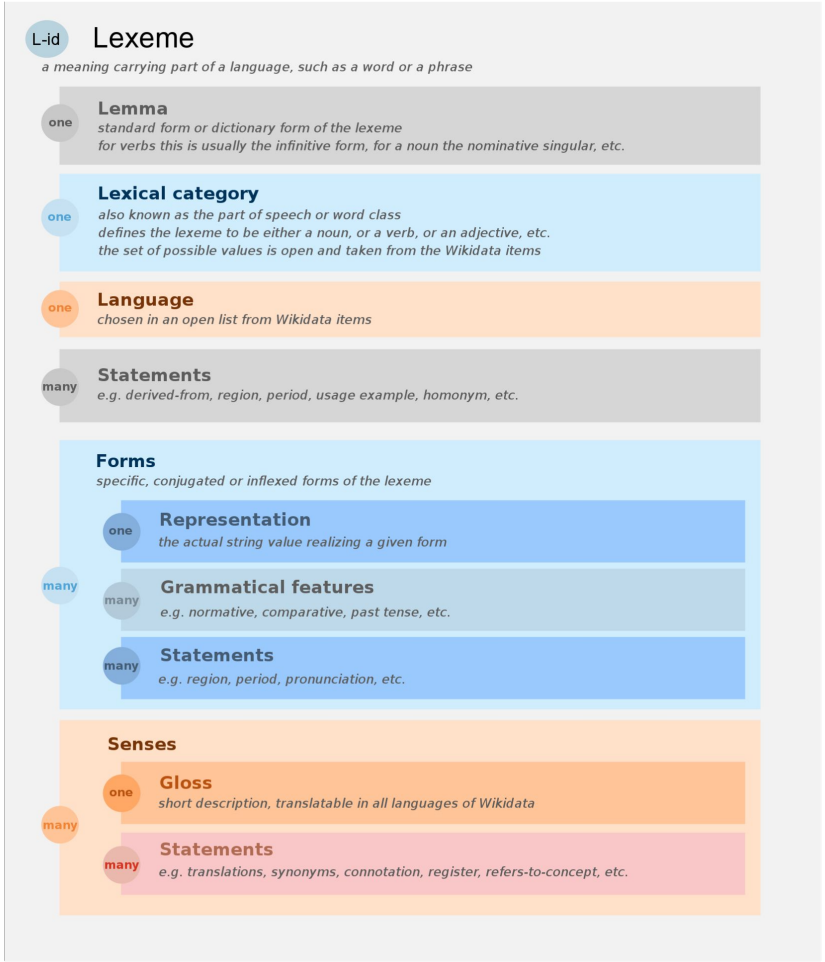
- Structured data = machine readable
- Can be reused by tools, research, dictionaries, translation services
- CC0 = open knowledge, can be reused by all
- Huge variety of languages, including underserved ones
- International community = more people to help

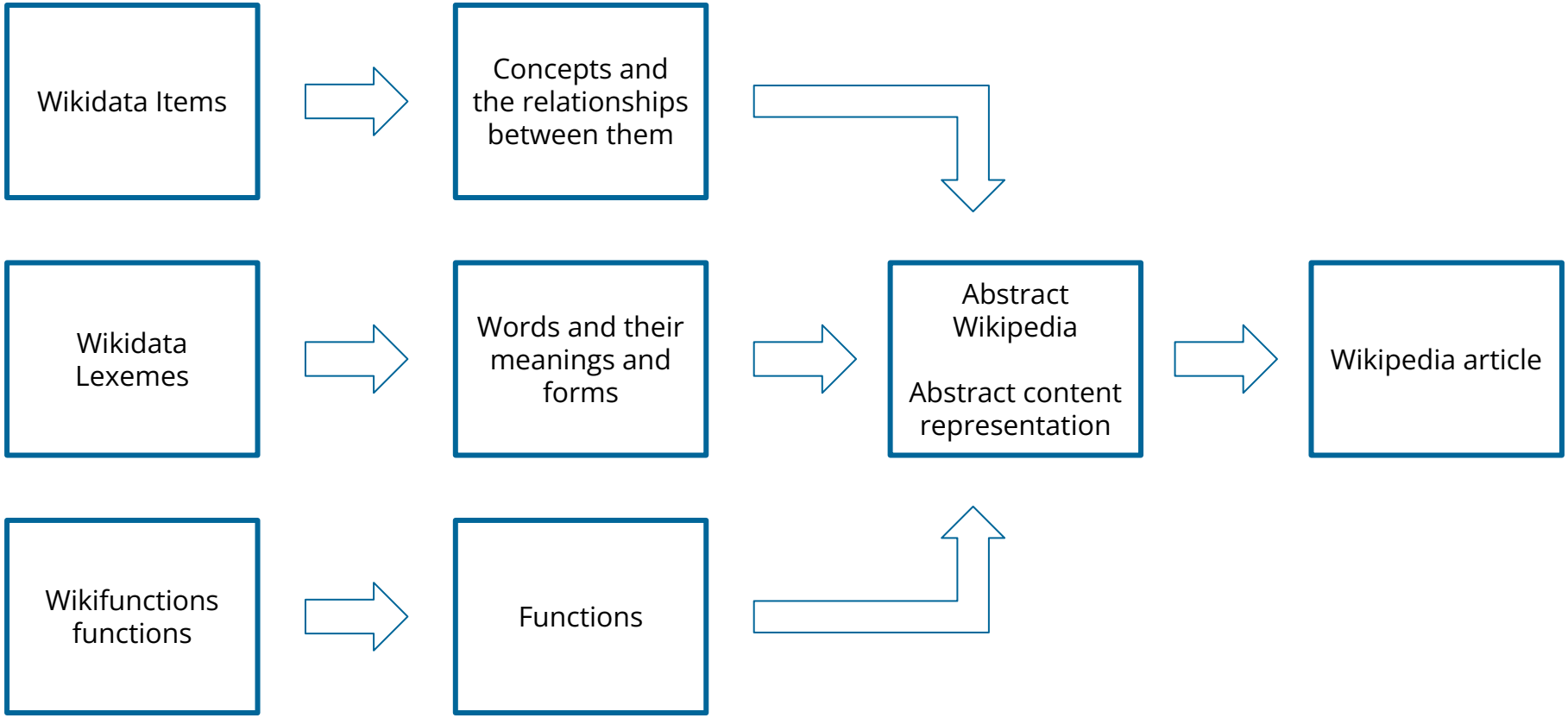
How is this different from other services?

- We're providing the background data to build anything on top of it
- We're doing much more than translation: we help machines understand languages
- We give access to the data in CC0
- We include all languages, not only the most profitable ones
- We empower people to contribute to the data

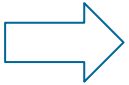
Conceptual development

- First data model proposal in 2013
- Heavily inspired by the Lemon model
- Several iteration with Wikidata community until 2015
- First software version deployed in 2018

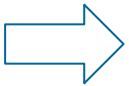



Wikidata Items
Concepts and the relationships between them
Wikidata Lexemes
Words and their meanings and forms
Abstract Wikipedia
Abstract content representation
Wikipedia article
Wikifunctions functions
Functions

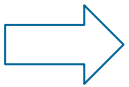
```
Article(
  content: [
        Instantiation(
              instance: San Francisco (Q62),
              class: Object_with_modifier_and_of(
                    object: center,
                    modifier: And_modifier(
                          conjuncts: [cultural, commercial, financial]
                    ),
                    of: Northern California (Q1066807)
              )
        ),
        Ranking(
              subject: San Francisco (Q62),
              rank: 4,
              object: city (Q515),
              by: population (Q1613416),
              local_constraint: California (Q99),
              after: [Los Angeles (Q65), San Diego (Q16552), San Jose (Q16553)]
        )
  ]
)
```



San Francisco is the cultural, commercial, and financial center of Northern California. It is the fourth-most populous city in California, after Los Angeles, San Diego and San Jose.



San Francisco ist das kulturelle, kommerzielle und finanzielle Zentrum Nordkaliforniens. Es ist, nach Los Angeles, San Diego und San Jose, die viertgrößte Stadt in Kalifornien.



...

Where are we going now?

- To enable truly meaningful applications we need more data (depth and breadth) and people to take care of it.
  - Focus languages: Hausa, Igbo, Dagbani, Malayalam, Bengali
- Abstract Wikipedia
- Exploring the secret magic sauce that is combining data about concepts and words in the same knowledge graph
- Expanding the Wikibase Ecosystem to explore how it can help create spaces for incubating new lex. data

Thank you

See you on Wikidata!

Email:
lydia.pintscher@wikimedia.de

Mastodon:
@nightrose@mastodon.online

Twitter:
@nightrose

Wikidata:
Q18016466

Bonus: Where is Wikidata as a whole going?

What are we focusing on now?

- Empower editors to increase data quality
- Facilitate equity in decision making
- Increase re-use for impact
- Strengthen underrepresented languages
- Enable Wikimedia Projects to share their workload

Empower editors to increase data quality

- Ensure that the content on Wikidata is of high quality for anyone who re-uses our data.
- Ensure that the socio-technical system is set up to help editors increase the quality of existing data and contribute high-quality new data.

Facilitate equity in decision making

- Ensure that fundamental decisions are made taking into account a diverse set of perspectives

Increase re-use for impact

- More people should benefit from the data Wikidata provides
- Our data is available for anyone to re-use. We want to especially support projects that are aligned with our mission and values and/or that give back to Wikidata.

Strengthen underrepresented languages

- More people should have access to technology that supports their language
- More people should have access to content in their language

Enable Wikimedia Projects to share their workload

- Wikimedia projects should be able to rely on Wikidata much more to provide content to their readers and maintain their content