Lexicographical data

An Introduction to Lexemes on Wikidata

Lydia Pintscher
Wikidata Portfolio Lead,
Wikimedia Deutschland

Mohammed Abdulai
Wikidata Community
Communication Manager,
Wikimedia Deutschland

This session is recorded: Please mute your microphone and camera when you're not speaking.

Wikidata Lexicodays
28-30 June 2024

Agenda

- Wikidata: A Quick Recap
- Words In Wikidata? How Does It Work?
- Why Is Lexicographical Data On Wikidata Important?
- What Could We Do With Lexemes?
- The State Of Lexicographical Data On Wikidata
- Useful Resources

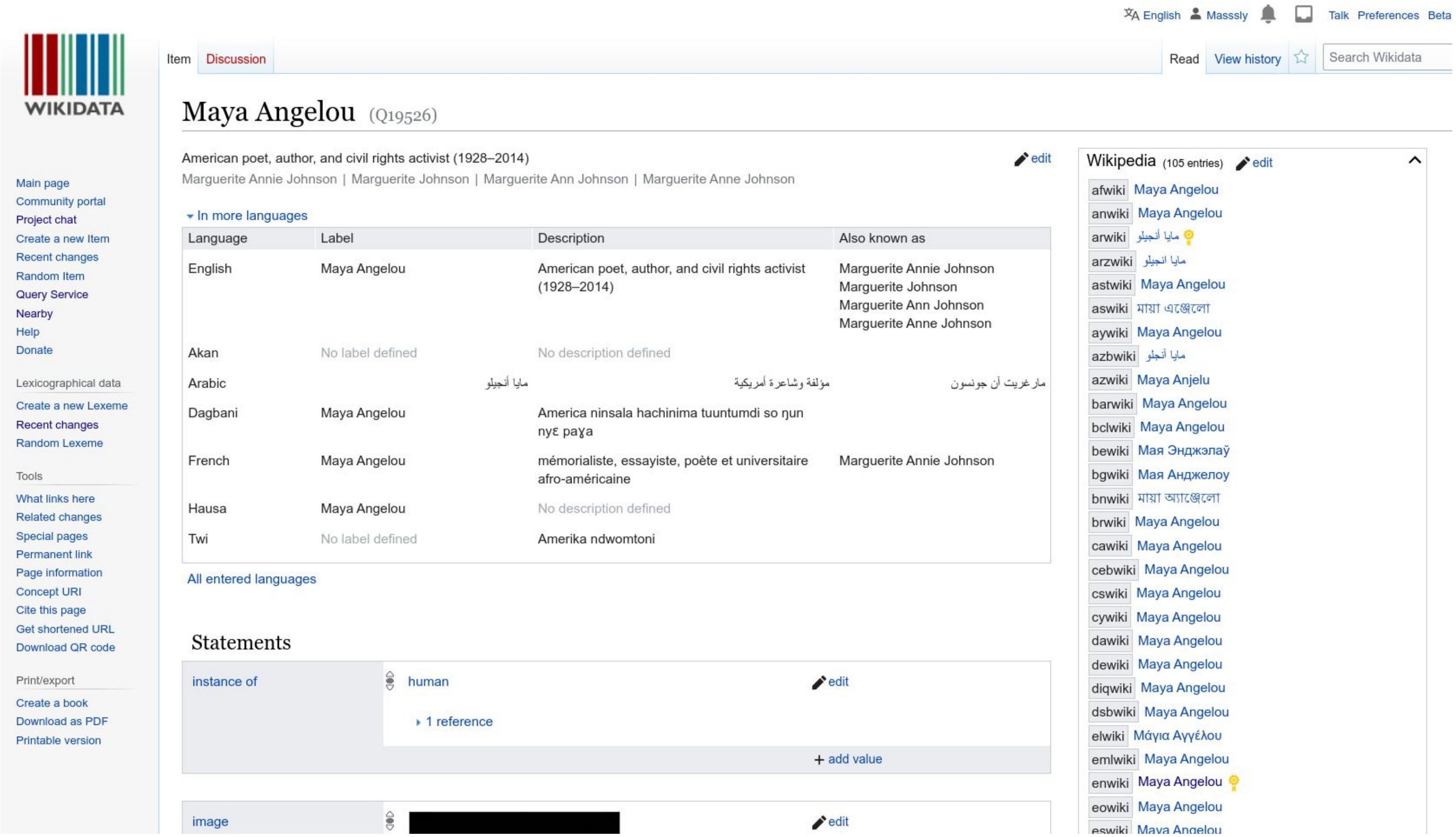
Wikidata: A quick recap

Wikidata
Lexicodays
28-30 June 2024

Wikidata: the basics

- A knowledge base
- Part of the Wikimedia projects
- Structured data
- Linked to other databases
- Multilingual
- Collaborative
- Released under public domain (CC0)
- Based on facts and references
- Made for humans and machines

Wikidata Item



Main page
Community portal
Project chat
Create a new Item
Recent changes
Random Item
Query Service
Nearby
Help
Donate

Lexicographical data

Create a new Lexeme
Recent changes
Random Lexeme

Tools

What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Cite this page
Get shortened URL
Download QR code

Print/export

Create a book
Download as PDF
Printable version

Maya Angelou (Q19526)

American poet, author, and civil rights activist (1928–2014) edit

Marguerite Annie Johnson | Marguerite Johnson | Marguerite Ann Johnson | Marguerite Anne Johnson

▾ In more languages

Language	Label	Description	Also known as
English	Maya Angelou	American poet, author, and civil rights activist (1928–2014)	Marguerite Annie Johnson Marguerite Johnson Marguerite Ann Johnson Marguerite Anne Johnson
Akan	No label defined	No description defined	
Arabic	مايا أنجيلو	مؤلفة وشاعرة أمريكية	مارغريت آن جونسون
Dagbani	Maya Angelou	America ninsala hachinima tuuntumdi so ŋun nyɛ paɣa	
French	Maya Angelou	mémorialiste, essayiste, poète et universitaire afro-américaine	Marguerite Annie Johnson
Hausa	Maya Angelou	No description defined	
Twi	No label defined	Amerika ndwomtoni	

All entered languages

Statements

instance of — human — edit

▸ 1 reference

+ add value

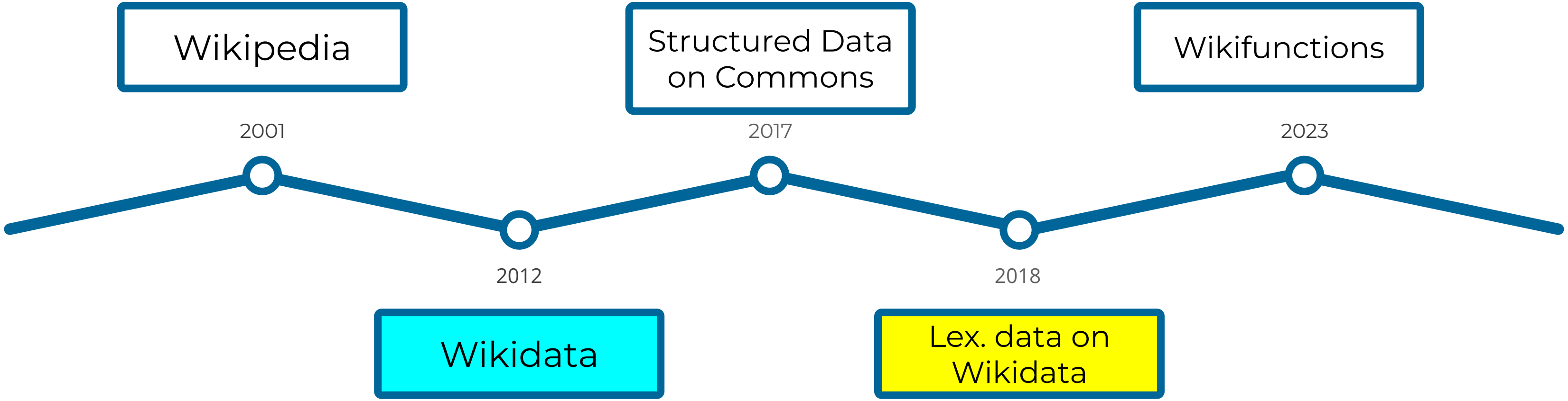
image — [redacted] — edit

Wikipedia (105 entries) edit

afwiki	Maya Angelou
anwiki	Maya Angelou
arwiki	مايا أنجيلو
arzwiki	مايا انجيلو
astwiki	Maya Angelou
aswiki	মায়া এঞ্জেলো
aywiki	Maya Angelou
azbwiki	مایا آنجلو
azwiki	Maya Anjelu
barwiki	Maya Angelou
bclwiki	Maya Angelou
bewiki	Мая Энджэлаў
bgwiki	Мая Анджелоу
bnwiki	মায়া অ্যাঞ্জেলো
brwiki	Maya Angelou
cawiki	Maya Angelou
cebwiki	Maya Angelou
cswiki	Maya Angelou
cywiki	Maya Angelou
dawiki	Maya Angelou
dewiki	Maya Angelou
diqwiki	Maya Angelou
dsbwiki	Maya Angelou
elwiki	Μάγια Αγγέλου
emlwiki	Maya Angelou
enwiki	Maya Angelou
eowiki	Maya Angelou
eswiki	Maya Angelou

¿Words in Wikidata: How does it work?

Wikidata
Lexicodays
28-30 June 2024



Wikipedia

2001

Wikidata

2012

Structured Data on Commons

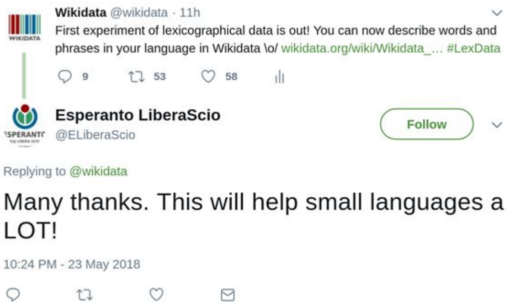
2017

Lex. data on Wikidata

2018

Wikifunctions

2023



Wikidata @wikidata · 11h

First experiment of lexicographical data is out! You can now describe words and phrases in your language in Wikidata \o/ wikidata.org/wiki/Wikidata_... #LexData

Esperanto LiberaScio
@ELiberaScio

Replying to @wikidata

Many thanks. This will help small languages a LOT!

10:24 PM - 23 May 2018

Wait, what’s the difference?

Concept “mouse”

- Species of mammal
- Taxon name
- Average size
- Picture
- Encyclopedia of Life ID

File:House mouse.jpg, public domain

Lexeme “mouse”

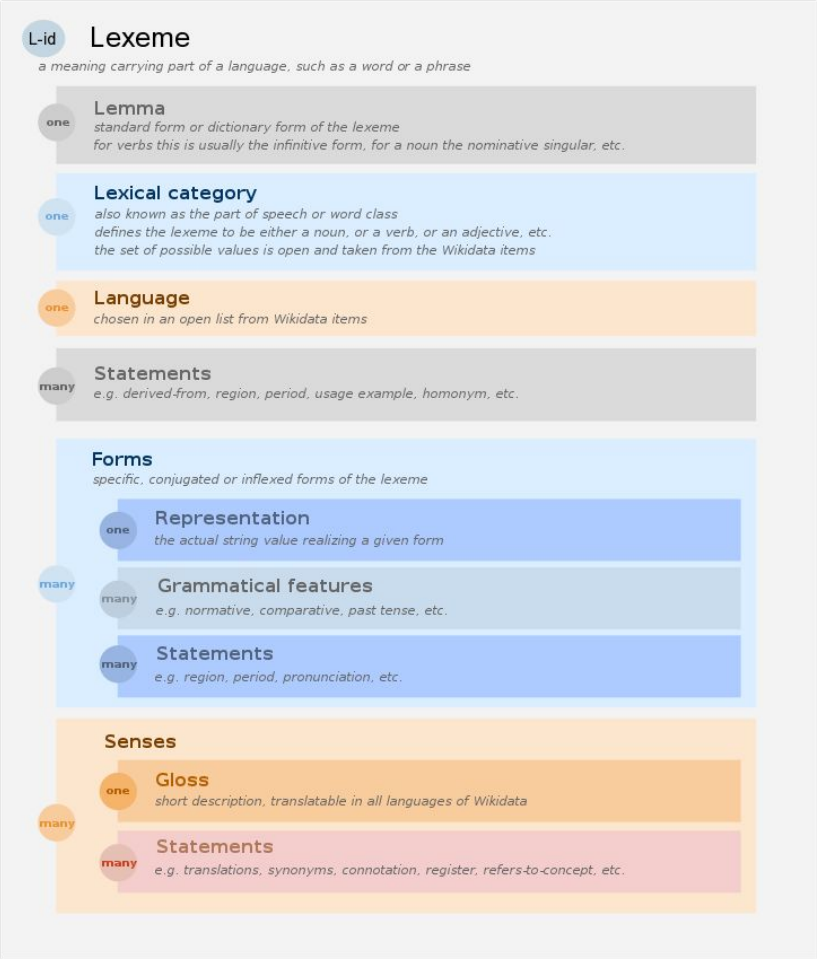
- Language: English
- Lexical category: noun
- Plural form: mice (irregular)
- Etymology: Proto-Germanic *mūs
- Senses: animal, computer device
- Translations: souris (fr), rato (pt)
- Audio pronunciation ▸

/maʊs/

Wikidata Lexicodays 28-30 June 2024

Glossary: Wikidata:Lexicographical_data/Glossary

Data model: MediaWiki:Extension:WikibaseLexeme/Data Model



L-id **Lexeme**

a meaning carrying part of a language, such as a word or a phrase

one **Lemma**
standard form or dictionary form of the lexeme
for verbs this is usually the infinitive form, for a noun the nominative singular, etc.

one **Lexical category**
also known as the part of speech or word class
defines the lexeme to be either a noun, or a verb, or an adjective, etc.
the set of possible values is open and taken from the Wikidata items

one **Language**
chosen in an open list from Wikidata items

many **Statements**
e.g. derived-from, region, period, usage example, homonym, etc.

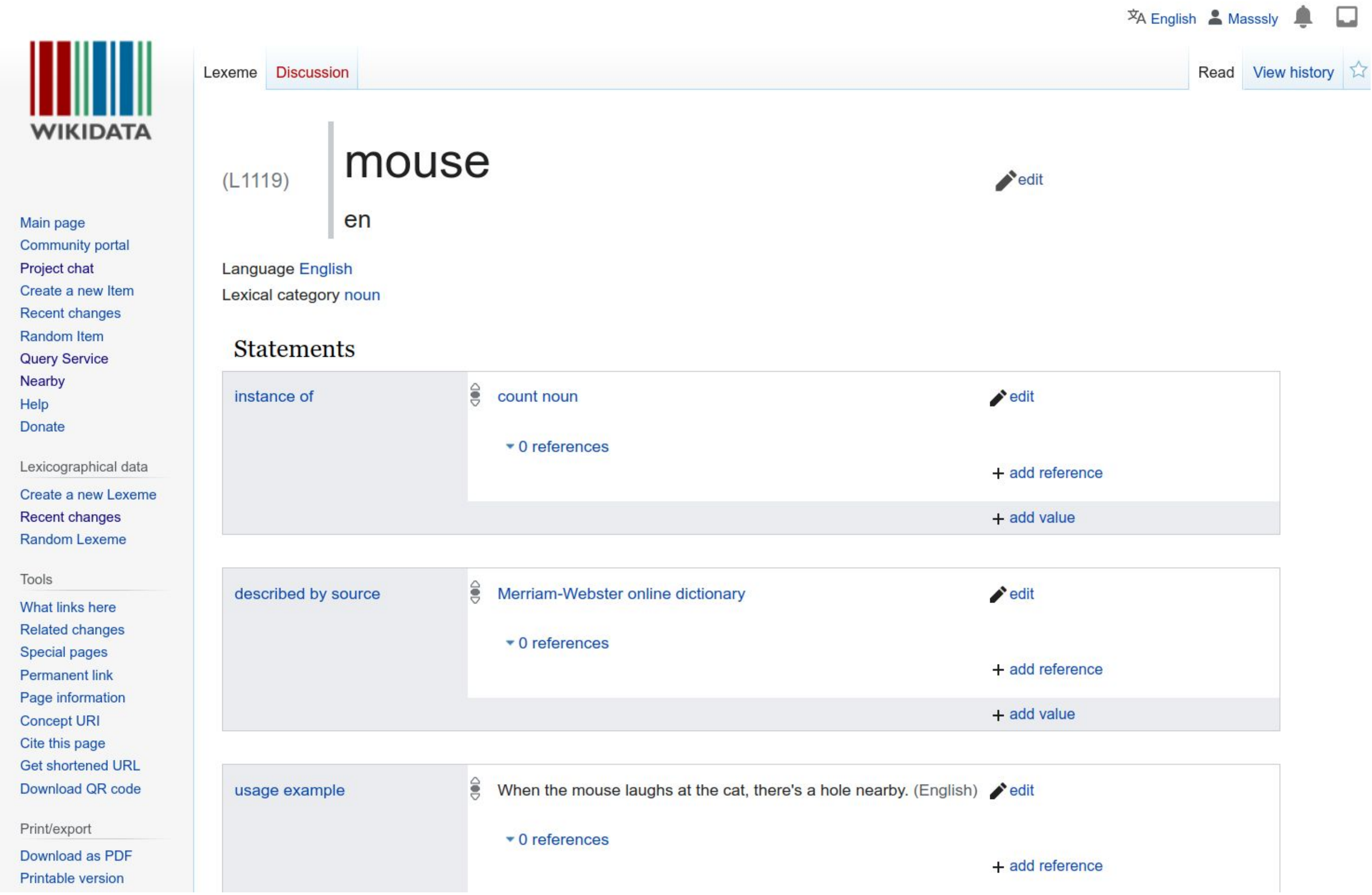
many **Forms**
specific, conjugated or inflexed forms of the lexeme

- one **Representation**
  the actual string value realizing a given form
- many **Grammatical features**
  e.g. normative, comparative, past tense, etc.
- many **Statements**
  e.g. region, period, pronunciation, etc.

many **Senses**

- one **Gloss**
  short description, translatable in all languages of Wikidata
- many **Statements**
  e.g. translations, synonyms, connotation, register, refers-to-concept, etc.

Wikidata Lexicodays
28-30 June 2024



(L1119) mouse

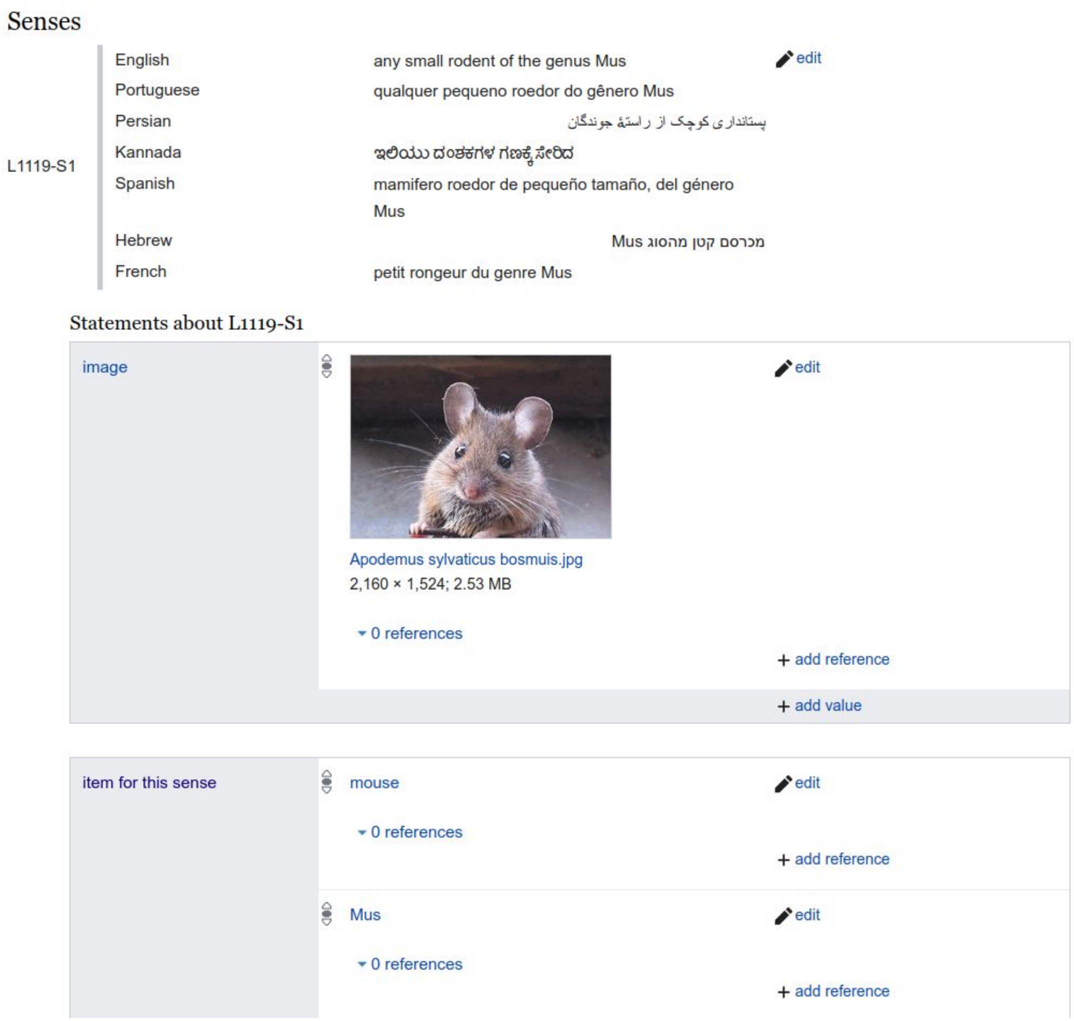
en

Language English

Lexical category noun

Statements

instance of	count noun 0 references
described by source	Merriam-Webster online dictionary 0 references
usage example	When the mouse laughs at the cat, there's a hole nearby. (English) 0 references



Senses

L1119-S1	English	any small rodent of the genus Mus	edit
	Portuguese	qualquer pequeno roedor do gênero Mus	
	Persian	پستانداری کوچک از راستهٔ جوندگان	
	Kannada	ಇಲಿಯು ದಂಶಕಗಳ ಗಣಕ್ಕೆ ಸೇರಿದ	
	Spanish	mamifero roedor de pequeño tamaño, del género Mus	
	Hebrew	מכרסם קטן מהסוג Mus	
	French	petit rongeur du genre Mus	

Statements about L1119-S1

image

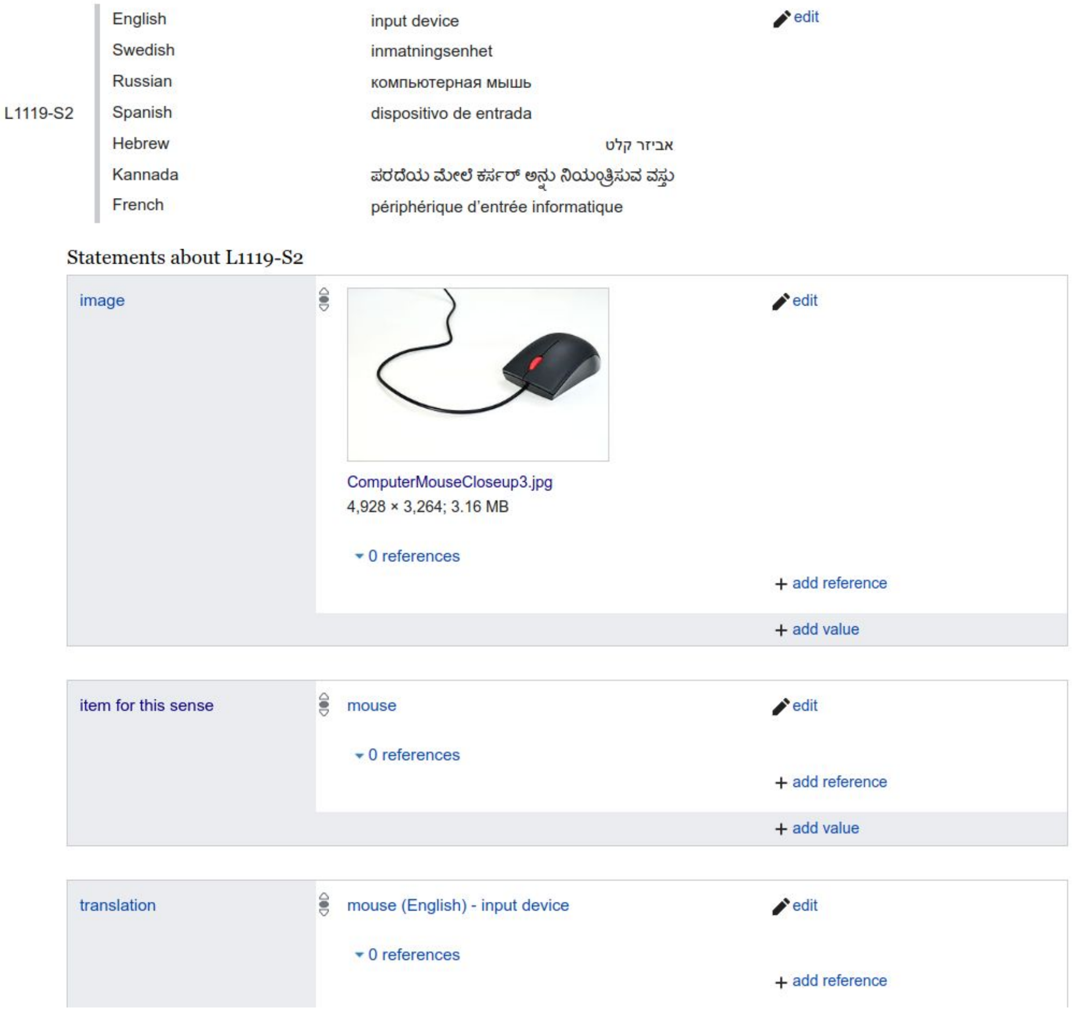
Apodemus sylvaticus bosmuis.jpg

2,160 × 1,524; 2.53 MB

item for this sense

mouse

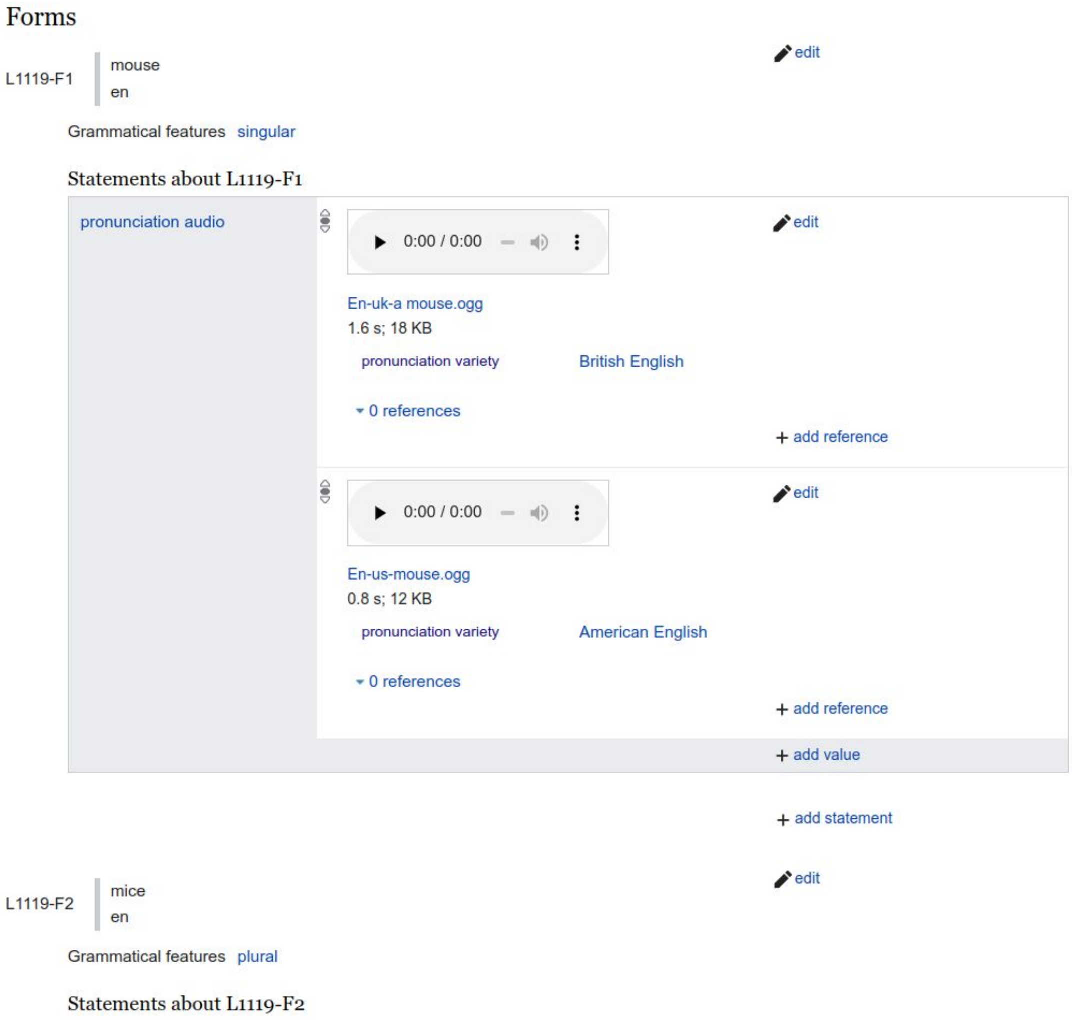
Mus



L1119-S2		
	English	input device
	Swedish	inmatningsenhet
	Russian	компьютерная мышь
	Spanish	dispositivo de entrada
	Hebrew	אביזר קלט
	Kannada	ಪರದೆಯ ಮೇಲೆ ಕರ್ಸರ್ ಅನ್ನು ನಿಯಂತ್ರಿಸುವ ವಸ್ತು
	French	périphérique d'entrée informatique

Statements about L1119-S2

image	ComputerMouseCloseup3.jpg 4,928 × 3,264; 3.16 MB 0 references	edit + add reference + add value
item for this sense	mouse 0 references	edit + add reference + add value
translation	mouse (English) - input device 0 references	edit + add reference



Forms

L1119-F1 | mouse
en

Grammatical features singular

Statements about L1119-F1

pronunciation audio

0:00 / 0:00

En-uk-a mouse.ogg
1.6 s; 18 KB

pronunciation variety British English

0 references

0:00 / 0:00

En-us-mouse.ogg
0.8 s; 12 KB

pronunciation variety American English

0 references

L1119-F2 | mice
en

Grammatical features plural

Statements about L1119-F2

Wikidata Lexicodays
28-30 June 2024

Why is Lexicographical Data on Wikidata Important?

Why is it interesting?

- 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

What’s the difference to Wiktionary?

- Wiktionary = plain text + templates, Wikidata = structured data
- Wikidata can be easily parsed and reused
- Wikidata works with Lexemes, Wiktionary combines different Lexemes in the same article
- Wikidata = CC0, Wiktionary = CC-BY SA
- Wikidata aims to support Wiktionaries (if they want to)

What’s the difference to 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

What could we do with Lexemes?

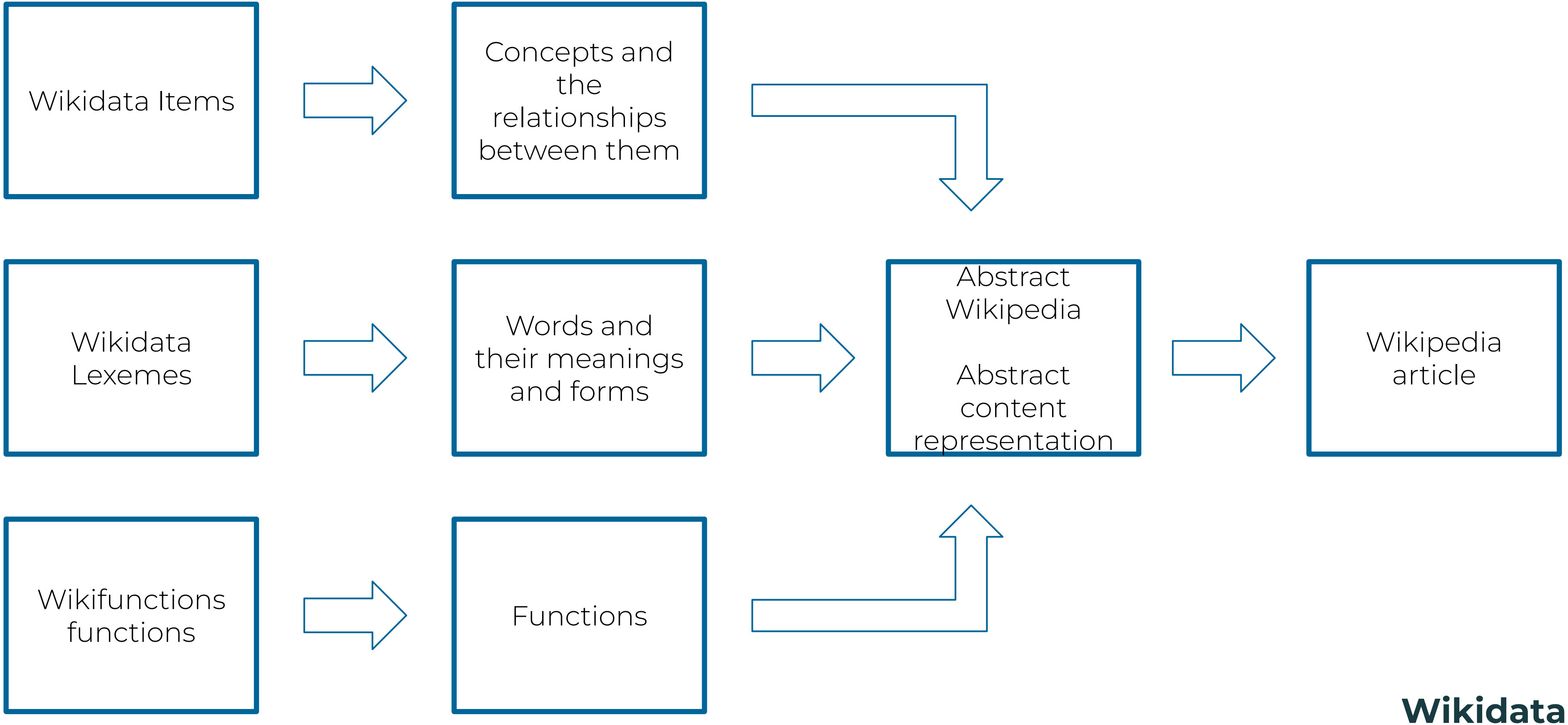
Support Wikimedia projects

- Provide structured data to be reused on pages
- Enable working together on the same data
- New tools to make contributing easier
- Text generation, eg. for Abstract Wikipedia

Abstract Wikipedia

- Human-readable text generation for Wikipedias
- …in different languages
- Based on available data

Components of Abstract Wikipedia



Wikidata Items

Concepts and the relationships between them

Wikidata Lexemes

Words and their meanings and forms

Abstract Wikipedia

Abstract content representation

Wikipedia article

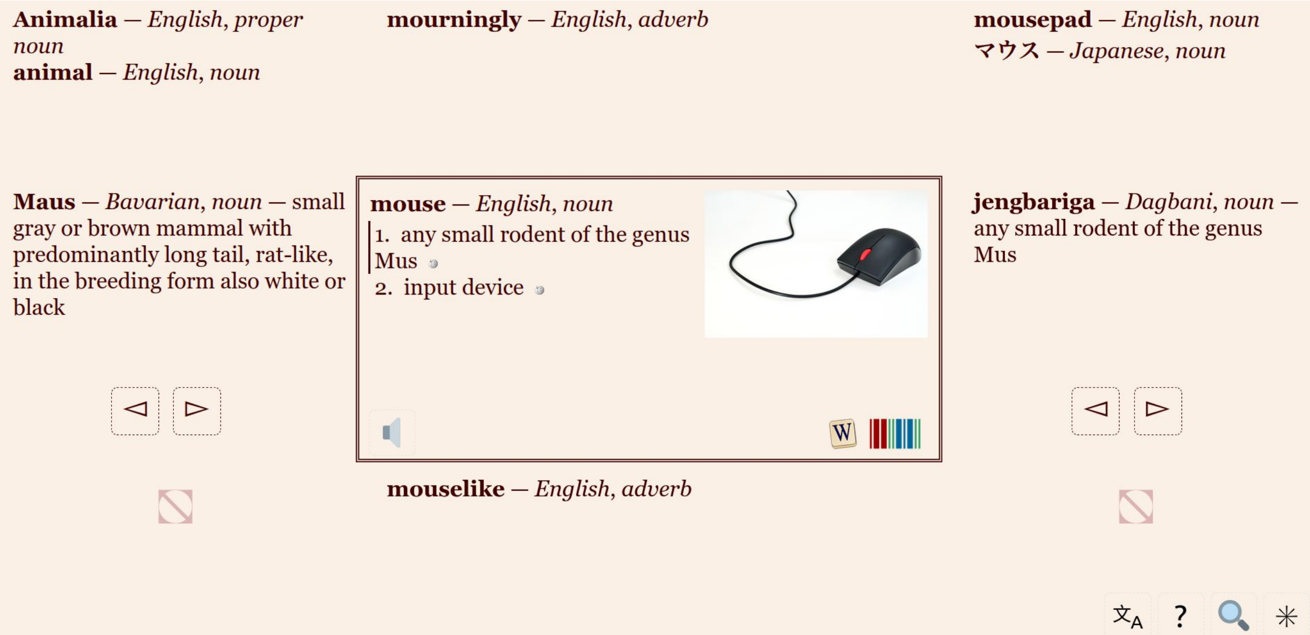
Wikifunctions functions

Functions

Dictionary applications

- Looking up definitions and translations
- Special purpose dictionaries (rhyme, specific topics)
- Thesauri and synonym dictionaries
- Build translation tools (especially for underserved languages that don't have any yet)

The Surrounding Ocean



Animalia — *English, proper noun*
animal — *English, noun*

mourningly — *English, adverb*

mousepad — *English, noun*
マウス — *Japanese, noun*

Maus — *Bavarian, noun* — small gray or brown mammal with predominantly long tail, rat-like, in the breeding form also white or black

mouse — *English, noun*

1. any small rodent of the genus Mus
2. input device

jengbariga — *Dagbani, noun* — any small rodent of the genus Mus

mouselike — *English, adverb*

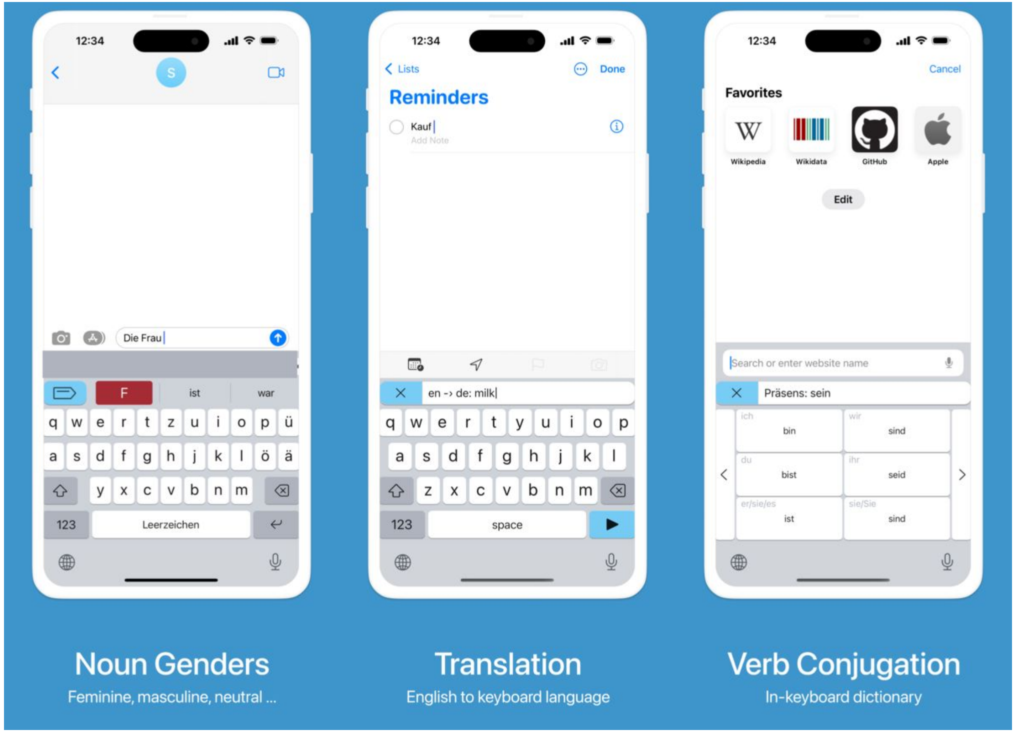
Explore words and their meanings, translations, and synonyms: d:Wikidata:The_Surrounding_Ocean

Wikidata Powered Language Keyboards

- Provide translations when typing in a second language
- Assist with Noun genders , verb conjugations & preposition cases
- Word annotation

Scribe Keyboard

- Keyboards for second language learners
- Currently in French, German, Italian, Portuguese, Russian, Spanish, Swedish, (more languages to be added)



ich bin	wir sind
du bist	ihr seid
er/sie/es ist	sie/Sie sind

Noun Genders
Feminine, masculine, neutral ...

Translation
English to keyboard language

Verb Conjugation
In-keyboard dictionary

https://github.com/scribe-org

Other Language learning tools

- Creating word lists and lessons
- Illustrating words
- Creating games and exercises

Research

- How do languages evolve over time, social class and more?
- Do classes of words change their meaning over time?
- Localizing words on maps

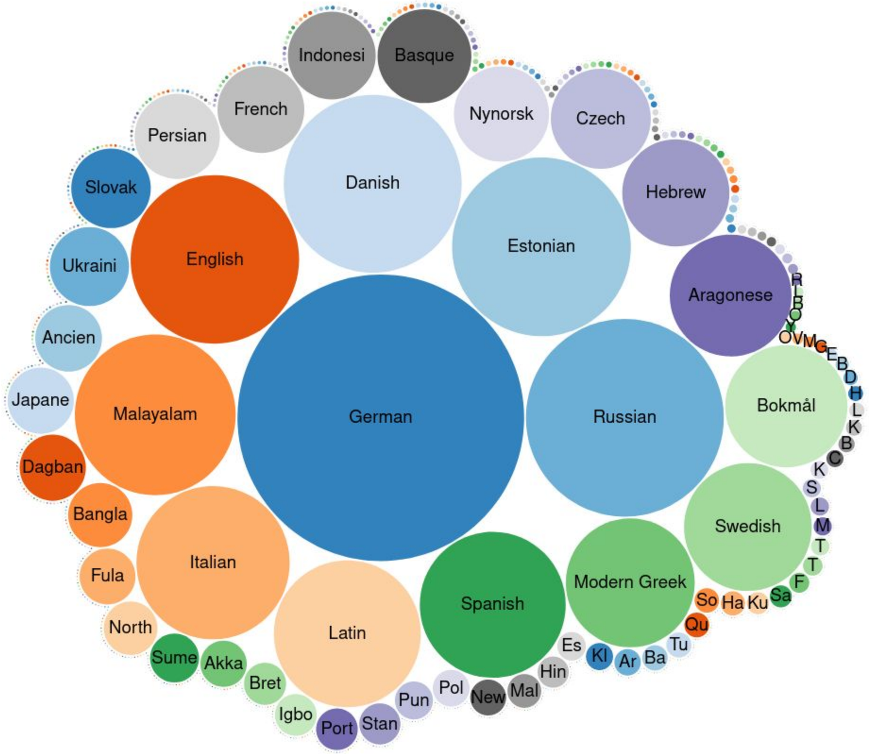
Text analysis

- Sentiment analysis
- Part of speech tagging
- Named entity recognition

The State of Lexicographical Data on Wikidata

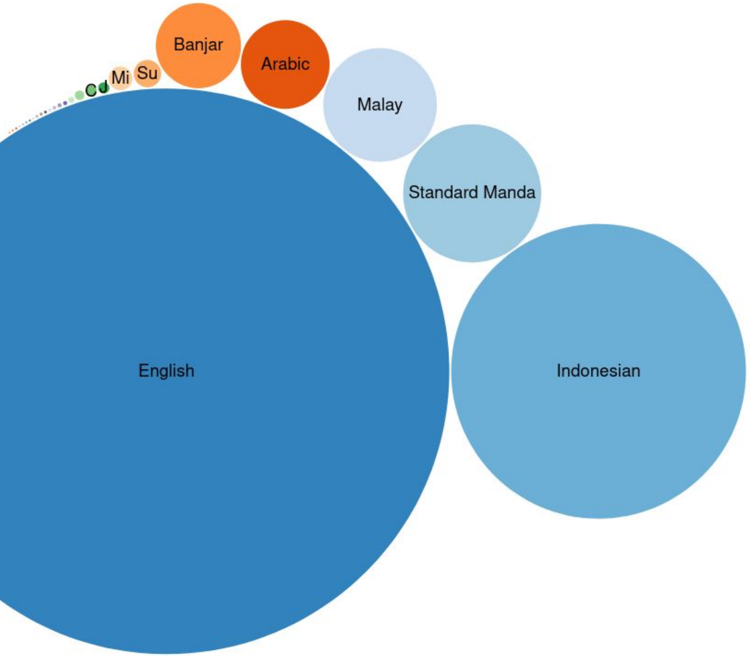
Wikidata
Lexicodays
28-30 June 2024

Distinct languages of Wikidata Lexemes



Try it: https://w.wiki/6RiP

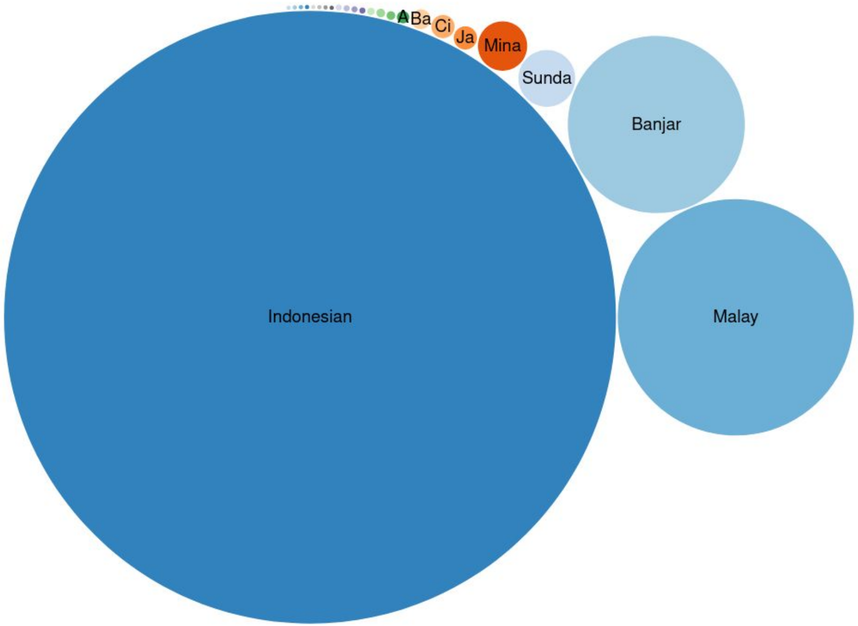
Distinct Indonesian languages of Wikidata Lexemes



Try it: https://w.wiki/AJ5U

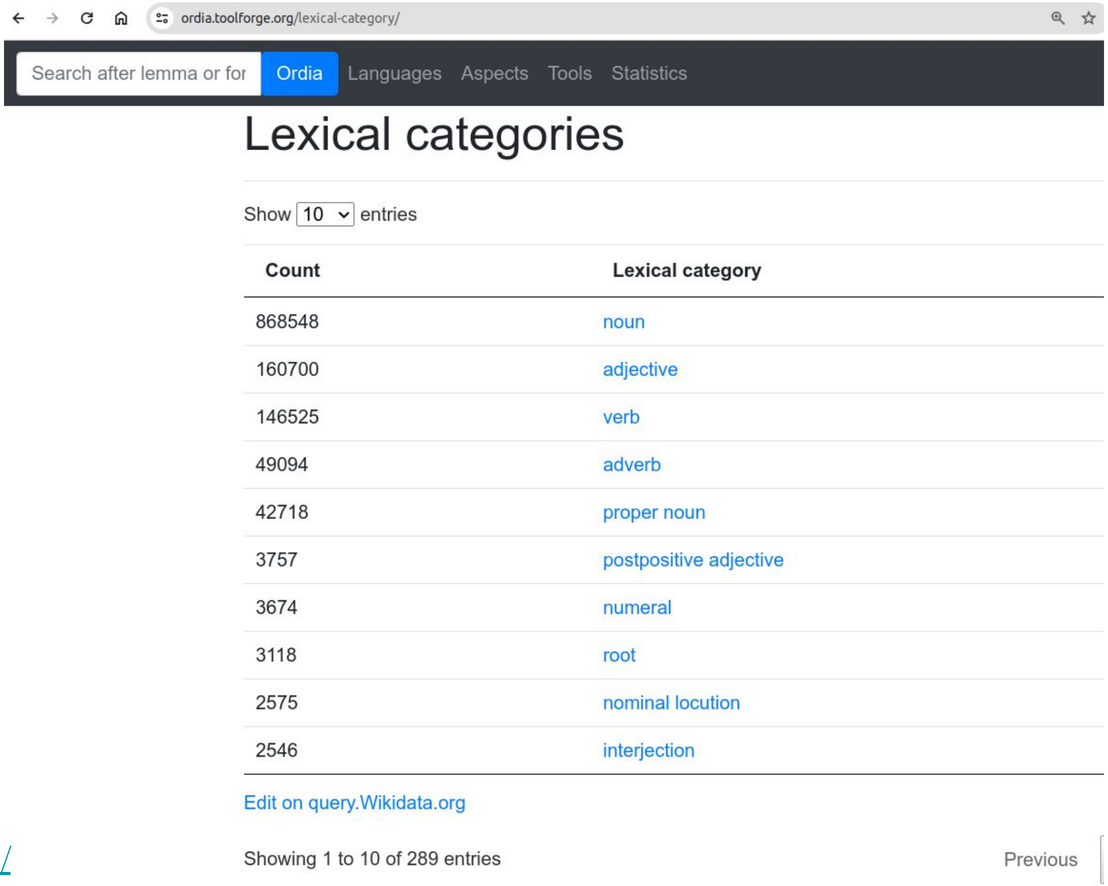
Wikidata
Lexicodays
28-30 June 2024

Indonesian languages Excl. En & Ar



Try it: https://w.wiki/AJ5u

Explore LexData coverage in more detail


ordia.toolforge.org/lexical-category/
Search after lemma or for
Ordia
Languages
Aspects
Tools
Statistics

Lexical categories

Show 10 entries

Count	Lexical category
868548	noun
160700	adjective
146525	verb
49094	adverb
42718	proper noun
3757	postpositive adjective
3674	numeral
3118	root
2575	nominal locution
2546	interjection

Edit on query.Wikidata.org

Showing 1 to 10 of 289 entries

Ordia:
https://ordia.toolforge.org/

Wikidata Lexicodays
28-30 June 2024

Explore LexData coverage in more detail

- Wikidata:Lexicographical coverage: d:Wikidata:Lexicographical_coverage
- Wikidata:Lexicographical data/Statistics: d:Wikidata:Lexicographical_data/Statistics
- Wikimedia Grafana: https://grafana.wikimedia.org/goto/INWiGyyIg?orgId=1

To enable truly meaningful applications we need more data (depth and breadth) and people to take care of it!

Useful Resources

Useful links

- Lexicographical Data Overview: Wikidata:Lexicographical_data
    - Try the existing tools Wikidata:Tools/Lexicographical_data
    - Suggest ideas of tools Wikidata:Lexicographical_data/Ideas_of_tools
    - Ideas & examples of queries Wikidata:Lexicographical_data/Ideas_of_queries
- Create and edit Lexemes Special:NewLexeme
- Suggest or discuss new properties Wikidata:Property_proposal/Lexemes
- Discuss with the community about how to model words
    - Wiki talk page: Wikidata_talk:Lexicographical_data
    - Telegram: https://bit.ly/4e8quK8

Thanks for your attention!

Get in touch with us:

Lydia Pintscher
lydia.pintscher@wikimedia.de
@nightrose everywhere

Mohammed Abdulai
mohammed.abdulai@wikimedia.de
@masssly everywhere

Credits

Wikidata
Lexicodays
28-30 June 2024