

# Lexicographical Data: words and phrases on Wikidata

Léa Lacroix  
Wikidata Labs, May 2019

# Table of contents

- Introduction
- Words in Wikidata: how does it work?
- Why is it interesting?
- Examples of reuse
- Frequently asked questions
- Editing Lexemes
- Tools to edit Lexemes
- Reusing Lexemes
- Resources
- Activities

# Wikidata: the basics



**WIKIDATA**

- 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



# Douglas Adams (Q42)

English writer and humorist

Douglas Noël Adams | Douglas Noel Adams | Douglas N. Adams | DNA

[edit](#)

▼ In more languages

Language	Label	Description	Also known as
English	Douglas Adams	English writer and humorist	Douglas Noël Adams Douglas Noel Adams Douglas N. Adams DNA
German	Douglas Adams	britischer Schriftsteller	Douglas Noël Adams Douglas Noel Adams
French	Douglas Adams	écrivain anglais de science-fiction	Douglas Noel Adams Douglas Noël Adams
Italian	Douglas Adams	scrittore inglese	Douglas Noel Adams Douglas N. Adams

All entered languages

Wikipedia (71 entries) [edit](#)

ar	دوغلاس آدمز
arz	دواجلس ادامز
ast	Douglas Adams
azb	داچلاس آدامز
az	Duqlas Adams
bar	Douglas Adams
be_x_old	Дуглас Адамз
be	Дуглас Адамс
bg	Дъглас Адамс
bn	ডগলাস অ্যাডামস
bs	Douglas Adams
ca	Douglas Adams
cs	Douglas Adams
cy	Douglas Adams
da	Douglas Adams
de	Douglas Adams
el	Ντάγκλας Αντάμης
en	Douglas Adams
eo	Douglas Adams
es	Douglas Adams
et	Douglas Adams
eu	Douglas Adams
fa	دواجلس آدامز
fi	Douglas Adams
fr	Douglas Adams
ga	Douglas Adams
gl	Douglas Adams
he	וּלְדָא סֶלְגַּט
hr	Douglas Adams

## Statements

instance of	<a href="#">human</a>	<a href="#">edit</a>
	▶ 2 references	

+ add value

image	<a href="#">Douglas adams portrait cropped.jpg</a>	<a href="#">edit</a>
	media legend	Portræt af Douglas Adams (Danish) Douglas Adams portrait. (English) Portrait de Douglas Adams. (French)



**Until 2018,  
Wikidata was  
only describing  
concepts**

**Now it also includes  
words!**

# Wait, what's the difference?

Concept “mouse”

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



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/

L-id

# Lexeme

**Lemma** - standard form or dictionary form of the lexeme

**Lexical category**

**Language**

**Statements** - e.g. derived-from, homonym, etc.

**Senses**

**Gloss** - short description

**Statements** - e.g. translations, synonyms, refers-to-concept, etc.

**Forms**

Representation

Grammatical features

**Statements** - e.g. region, period, pronunciation, etc.

More info:  
[mw:Extension:Wikibase Lexeme/Data Model](#)

(L407)

# casa

edit

pt

Language [Portuguese](#)

Lexical category noun

grammatical gender

feminine

edit

[▼ 0 references](#)[+ add reference](#)[+ add value](#)

derived from

casa

edit

[▼ 0 references](#)[+ add reference](#)[+ add value](#)

## Senses

L407-S1	English	building, habitation	 edit
	Portuguese	prédio	

### Statements about L407-S1

#### translation

 house (structure built or serving as an abode of human beings)  edit

▼ 0 references

+ add reference

+ add value

## Forms

L407-F1	casa	 edit
	pt	

Grammatical features [plural](#)

### Statements about L407-F1

#### IPA transcription

 /'ka.zas/  edit

language of work or name

Brazilian Portuguese

▼ 0 references

+ add reference

# Why is it interesting?

- Structured data = machine readable
- Plain text = only for humans, need to build complex algorithms
- Can be reused by tools, research, dictionaries, translation services
- CC0 = open knowledge, can be reused by all
- Huge variety of languages, including undeserved ones
- International community = more people to help

# Examples of reuse: question answering algorithms

 QAnswer  Go About FAQ    

Confidence :  87 % SPARQL LIST DID YOU MEAN DIRECT ANSWER

Is this the right answer ?  Yes  No

*dragons*

**dragons** + 

# Examples of reuse: language-learning games

## DerDieDas

Practice your German articles with Wikidata

What is the correct article?

- der
- die
- das
- I don't know

## Funktion

Submit



Round: 1/10



Points: 0

Der Die Das <http://auregann.fr/derdiedas/>

# Other uses in the future

- Support for Wiktionary, Wikisource & other Wikimedia projects
- Research
- Text analysis, text mining
- Translation tools
- New generation of dictionaries
- Language-learning tools

# A few numbers

- **45400** Lexemes in **340** languages and dialects
- **42** Lexemes in Portuguese
- **12000** Senses (definitions)
- More statistics: <https://tools.wmflabs.org/ordia/>

# Frequently asked questions

## What's the difference with Wiktionary?

- Wiktionary = plain text + templates, Wikidata = structured data
- Wikidata can be easily parsed and reused
- Wikidata works with Lexemes, Wiktionary combines Lexemes
- Wiktionary may have extra info (examples, quotes...)
- Wikidata aims to support Wiktionaries (if they want to)

# Frequently asked questions

## How to enter dialects or scripts?

- Use mis-x-Q... as language of the lemma

(L12607)	صفر	صِفْرٌ
ar		ar-x-Q775724

Language Arabic

Lexical category numeral

# Frequently asked questions

## How to search for Lexemes?

- Type “L:word” in the search box and press Enter
- Works with lemmas and Forms
- No suggestion appears, don’t worry

### Search results

To search for Wikidata items by their title on a given site

L:casa

**Advanced parameters:**

**Search in:** (Main) X (Property) X

casa (L407)  
Portuguese, noun  
6 statements, 2 forms - 10:21, 18 May 2019

casa (L350)  
Italian, noun  
1 statement, 2 forms - 10:21, 18 May 2019

casa (L438)  
Latin, noun  
0 statements, 0 forms - 10:21, 18 May 2019

# Frequently asked questions

## Is it possible to query Lexemes?

- Yes! The Query Service supports Lexemes
- All words in Portuguese <https://w.wiki/45n>
- Longest words in English <https://w.wiki/45o>
- More queries & ideas:  
[https://www.wikidata.org/wiki/Wikidata:Lexicographical\\_data/Ideas\\_of\\_queries](https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Ideas_of_queries)
- Get help: [https://www.wikidata.org/wiki/Wikidata:Request\\_a\\_query](https://www.wikidata.org/wiki/Wikidata:Request_a_query)

# Frequently asked questions

## Does QuickStatements work for Lexemes?

- Not really :(
- Possible: add new statements on the Lexeme level for existing Lexemes
- Not possible: create new Lexemes, Forms or Senses



# **Any questions so far?**

# Editing Lexemes

- Check if the Lexeme exists (L:singen)
- Find an example <https://www.wikidata.org/wiki/Lexeme:L2058>
- Create a new Lexeme <https://www.wikidata.org/wiki/Special>NewLexeme>
- Add statements (auxiliary verb)
- Add a Sense
- Add translation (needs another Sense)
- Add a Form (present, first-person singular)

# A few tips for editing

- Look at what already exists before creating
- Find an example for inspiration
- Look for the most used properties in Ordia  
<https://tools.wmflabs.org/ordia/property/>
- Try, don't be afraid of doing wrong :)
- In doubt: ask the other editors for help ([LexData talk page](#), [Telegram](#))

# **Everything is new!**

- Started from scratch in May 2018
- Everything is to create
- Still plenty of questions, ways to models
- Maybe new properties to be created
- Not everything works perfectly (eg mobile interface)
- Some documentation needs improvement

# Tools to edit Lexemes

- Wikidata Lexeme Forms, to create Lexemes and generate a lot of Forms  
<https://tools.wmflabs.org/lexeme-forms/>
- Possibility to create new templates (eg. conjugations, declinations)

## English verb

### present

They  every day.

### third-person singular

He  every day.

### simple past

He  every day last week.

### present participle

They are  right now.

### past participle

We have  for hours.

Create

Advanced

# Tools to edit Lexemes

- Wikidata Senses, to add Senses to existing Lexemes  
<https://tools.wmflabs.org/lexeme-senses/>
- Add Senses for Portuguese words in Portuguese:  
<https://tools.wmflabs.org/lexeme-senses/add/pt>

The screenshot shows a web-based form for editing a lexeme. At the top, there are two yellow buttons: 'Ordia' and 'Wikidata'. Below them, the word 'zebro' is displayed in large, bold, black font. Underneath the word, it says 'pt · noun · L5451'. A placeholder text 'Add the first Sense to this Lemma!' is followed by a long, empty input field. At the bottom, there are two buttons: 'Give me another lemma' on the left and 'Add' on the right.

Ordia    Wikidata

zebro

pt · noun · L5451

Add the first Sense to this Lemma!

Give me another lemma    Add

# Tools to edit Lexemes

- Ordia text-to-Lexeme transforms a text into Lexemes and helps with creation <https://tools.wmflabs.org/ordia/text-to-lexemes>

## Text to lexemes

A Barata diz que tem sete saias de filó  
É mentira da barata, ela tem é uma só  
Ah ra ra, iá ro ró, ela tem é uma só !



ro					
ró					
aias					
ete					
ó					
em	tem	ter	verb	third-person singular // present indicative	L41087-S1
ima					
:					
é	é	ser	verb	third-person singular // present indicative	L39470-S1
é	é	ser	verb	third-person singular // present indicative	L39470-S2

# Tools to edit Lexemes

- Lingua Libre, record words in your language  
[https://lingualibre.fr/wiki/LinguaLibre:Main\\_Page](https://lingualibre.fr/wiki/LinguaLibre:Main_Page)
- Generate lists of words from Lexemes  
[https://lingualibre.fr/wiki/Help:Create\\_your\\_own\\_lists](https://lingualibre.fr/wiki/Help:Create_your_own_lists)
- Automatically adds the pronunciation as a statement



# Tools to reuse Lexemes

- The Query Service
  - Lexemes don't appear in the documentation... yet ^^
- The API

wbladdform

wbladdsense

wbleditformelements

wbleditsenseelements

wblinktitles

wblmergelexemes

wblremoveform

wblremovesense

# Helpful pages

- Create and edit Lexemes [Special:NewLexeme](#)
- Suggest or discuss new properties [Wikidata:Property\\_proposal/Lexemes](#)
- Discuss with the community about how to model words &
- Report any bug or wish for the future [Wikidata\\_talk:Lexicographical\\_data](#)
- Try the existing tools [Wikidata:Tools/Lexicographical\\_data](#)
- Suggest ideas of tools [Wikidata:Lexicographical\\_data/Ideas\\_of\\_tools](#)
- Ideas & examples of queries  
[https://www.wikidata.org/wiki/Wikidata:Lexicographical\\_data/Ideas\\_of\\_queries](https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Ideas_of_queries)

# More questions?

# Activities

- Add Senses in Portuguese  
<https://tools.wmflabs.org/lexeme-senses/add/pt>
- Create more Lexemes in Portuguese
  - Using text-to-Lexemes  
<https://tools.wmflabs.org/ordia/text-to-lexemes>

# Activities

- Create Portuguese templates in Wikidata Lexeme Forms
  - Create the page for Portuguese  
[https://www.wikidata.org/wiki/Wikidata:Wikidata\\_Lexeme\\_Forms#Language\\_support](https://www.wikidata.org/wiki/Wikidata:Wikidata_Lexeme_Forms#Language_support)
  - Help translating the interface
  - Add new templates (example:  
[https://www.wikidata.org/wiki/Wikidata:Wikidata\\_Lexeme\\_Forms/French](https://www.wikidata.org/wiki/Wikidata:Wikidata_Lexeme_Forms/French))
  - Ping Lucas (user:Lucas Werkmeister) to include the templates in the tool (or for any question)

# Activities

- Once templates are ready: create Forms in Portuguese
- Developers: create a tool or a game using lexicographical data (see also: [ideas of tools](#))

# Thank you!

Léa Lacroix

User: Lea Lacroix (WMDE)

[lea.lacroix@wikimedia.de](mailto:lea.lacroix@wikimedia.de)