Lexicographical Data has arrived on Wikidata



Léa Lacroix Celtic Knot Conference, July 2018

Vou can now enter Data about words and phrases Freely reusable Structured to be machine-readable Improved by a multilingual community that can be used by:

WiktionariesWikidataStudentsNGOsWikisourceScientistsPublic institutionsStart-upsData-journalists



Workload sharing and new ways to contribute to Wiktionary

- Working together on the same data (if wanted!)
- New tools to make contributing easier and open it up to new contributor groups

Potential users: Wiktionary, Wikidata Game

Dictionary applications and more

- 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)

Potential users: Leo, Apertium

Language learning tools

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

Potential users: Parley, Duolingo

Research

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

Potential users: The Rosetta Project

Text analysis

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

Potential users: TextRazor, Wikisource



L-id Lexeme

Lemma - standard form or dictionary form of the lexeme

Lexical category

Language

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

Forms	5
	Representation
	Grammatical features
	Statements - e.g. region, period, pronunciation, etc.
Sense	25

Gloss - short description

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

More info: <u>mw:Extension:WikibaseLexeme/Data Model</u>

(L4729) Aberystwyth

/ edit

Language Welsh Lexical Category toponym

Statements



(L2285)

cwrw _{cy}

Language Welsh Lexical Category noun

Statements

Forms

L2285-F1 cyrfau cy Grammatical features plural Statements about L2285-F1

L2285-F2 cwrw cy Grammatical features singular

Statements about L2285-F2

described by source	Q23705356		/ edit	
	volume	1		
	page(s)	221		
	publication date	1880		
	full work available at	https://ru.wikisource.org /wiki/TCД2/Вода		
	- 0 references			
			+ add reference	
			+ add value	

grammatical gender	feminine	dit dit
	+ 0 references	 add reference
		+ add value

derived from	🛢 вода	/ edit
	 0 references 	
		+ add reference
		+ add value

	d	d		•	2	÷		-		÷
 9	4	ч.	-	1	9	۲	-			۰.

add reference
 add value

Forms

L189-F1 BODU ru Pedit ru Grammatical features singular, genitive case Statements about L189-F1 trapeliteration È votv Padit

transiteration	5 VOUY	edit
	✓ 0 references	
		+ add reference
		+ add value

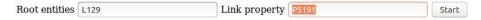


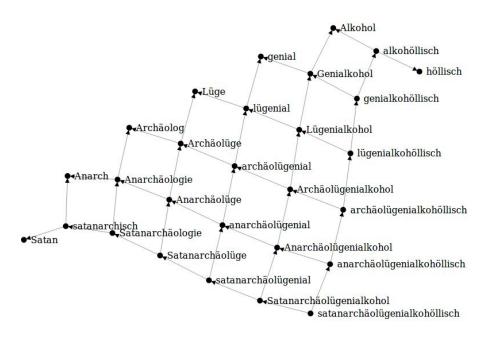
1	<u>89</u>	: в	ОД	a,	ru,	noun	
---	-----------	-----	----	----	-----	------	--



- Create and edit Lexemes **Special:NewLexeme**
- Try if it's working with your language
- Suggest or discuss new properties <u>Wikidata:Property_proposal/Lexemes</u>
- Discuss with the community about how to model words
- Report any bug or wish for the future <u>Wikidata_talk:Lexicographical_data</u>
- Try the existing tools <u>Wikidata:Tools/Lexicographical_data</u>
- Suggest ideas of tools <u>Wikidata:Lexicographical_data/Ideas_of_tools</u>

Wikidata Lexeme Graph Builder







- Better search features
- Improvements on the interface
- Senses
- Possibility to build queries with the data
- More tools
- More editors, more diversity in the languages!



During the first month after deployment, there have been:

5000 Lexemes created

By around **120** editors

In more than **100** different languages

100+ discussions and suggestions on wiki and Phabricator

5 external tools created



Follow

#Wikidata has just joined the world of #Lexicography! Not since the first edition of the @OED has crowdsourcing and dictionaries come together quite so tightly.

Wikidata @wikidata

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

5:24 PM - 23 May 2018



Suivre

Small tweet for an important first step. @Wikidata's foray into lexicographical data has huge potential to liberate humanity's access to our shared heritage of languages.

Erik Moeller @xirzon · 44 min

The possible applications of an open, collaborative repository of lexicographical data in all languages include written and visual dictionaries, translation tools, language and vocabulary learning helpers, machine learning applications, and much more.

Traduire le Tweet

17 0 1 M 0 1

Erik Moeller @xirzon · 38 min

Aside from the fact that this data is free to use for anyone, it also is not subject to the same market pressures. If there is a community that creates #LexData in Swahili or Kannada, developers can just as easily use it to build applications as they would French or English.



Replying to @wikidata

Many thanks. This will help small languages a I OTI

afrocrowd.org afro CROWD @afroCROWDit

Replying to @wikidata

Good stuff! Thanks, we needed it!

John Samuel @isamwrites



Follow



Suivre

This is a significant development, and wellsuited to the Wiki approach

Denny Vrandečić @vrandezo

Wikidata launched a few hours ago their first experimental support for lexicographic data - i.e. it can now store words and knowledge about words. Within a few hours, more than 600 entries in 38 languages were created. Congratulations to the ...

> Follow \sim

Great choice. Lexeme: L1 "mother" #LexData #wikidata wikidata.org/wiki/Lexeme:L1



Wikidata @wikidata · 10h

Replying to @AlbinPCLarsson @fnielsen

Thanks for reporting. We're working on making it more understandable.





Albin Larsson @AlbinPCLarsson · 10h You are awesome and are doing a great job! :-)

Congrats on shipping this new feature!





Thanks for your attention :)