



Lexicographical Data has arrived on Wikidata

Léa Lacroix
Celtic Knot Conference, July 2018

You 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:

Wiktionaries

Wikidata

Students

NGOs

Wikisource

Scientists

Public institutions

Start-ups

Data-journalists

**What can we do
with it?**



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

How is it working?



L-id

Lexeme

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

Lexical category

Language

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

Forms

Representation

Grammatical features

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

Senses

Gloss - *short description*

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

More info: [mw:Extension:WikibaseLexeme/Data Model](https://www.wikidata.org/wiki/Extension:WikibaseLexeme/Data_Model)

(L4729)

Aberystwyth

cy

 edit

Language [Welsh](#)

Lexical Category [toponym](#)

Statements

| | |
|-------------|---|
| compound of |  Ystwyth  edit |
| | series ordinal 2 ▼ 0 references + add reference |
| |  aber  edit |
| | series ordinal 1 ▼ 0 references + add reference + add value |

(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

**What can I
already do?**



- Create and edit Lexemes <Special:NewLexeme>
- Try if it's working with your language
- 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

**What will
come next?**



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

A few numbers



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

 **Liam Wyatt** @Wittylama 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_...](http://wikidata.org/wiki/Wikidata_...) #LexData

5:24 PM - 23 May 2018

 **Erik Moeller** @xirzon 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
1 1 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.

 **Esperanto LiberaScio** @ELiberaScio Follow

Replying to @wikidata
Many thanks. This will help small languages a LOT!

 **afrocrowd.org** @afroCROWDit

Replying to @wikidata
Good stuff! Thanks, we needed it!

JS **John Samuel** @jsamwrites Follow

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.

1 4

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

Congrats on shipping this new feature!

 **Dan Brickley** @danbri Suivre

This is a significant development, and well-suited 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

Questions?

Léa Lacroix

User: Lea Lacroix (WMDE)

lea.lacroix@wikimedia.de

Thanks for your attention :)