Wikidata for Wiktionary

Let’s get ready for lexicographical data!

2016, September 12th

Authors: Léa Lacroix, Lydia Pintscher, Daniel Kinzler, Denny Vrandečić

What is Wikidata?

- A knowledge base
- Storing structured information
- Free license
- One dataset, hundreds of languages
- Supporting Wikimedia projects
- Providing data to third parties

What is Wiktionary?

- A dictionary
- Free license
- Different datasets in several language wikis
- Goal: providing dictionary entries of all words in all languages

Wikidata for Wiktionary: what we want to do

- Support Wiktionary editors and content
- Make lexicographical data available in a structured and machine-readable way, that allows the information to be shown in multiple languages with no additional effort. It allows for better search, easier re-use, and new tools.

Why ? How this will be useful for Wiktionaries?

- Enable editors to collaborate across Wiktionaries more easily
- Increase the number of editors and visibility of languages
- Improve the quality of data by increasing the number of people working on it
- Provide the groundwork for new tools for editors

Layout of a typical Wiktionary entry

Wiktionary projects maintained in different languages have somewhat different structures. In particular, the role of etymology in the structure seems to vary. The placement of the pronunciation also varies, depending on whether the pronunciation is the same for all variants, or not. Translations are cross-linked to individual senses, and generally not associated with them structurally.

A Wiktionary page may have this structure:

- Page ("word")
  - Language
    - **Morphological category** (verb, noun, male, female...)
      - Headword line
      - **Pronunciation**
      - **Etymology**
      - Definitions (senses)
        - **Sense 1**
        - …
      - **Translations**
        - **Sense 1**
        - **...**
    - Morphological category
      - …
  - Language
    - …

Layout of a typical Wiktionary entry

or like this…

- Page (“word”)
  - Language
    - **Etymology**
      - Morphological category (verb, noun, male, female...)
        - Headword line
        - **Pronunciation**
        - Definitions (senses)
          - Sense 1
          - …
        - Translations
          - Sense 1
          - ...
      - Morphological category
      - …
  - Language
    - …

Layout of a typical Wiktionary entry

or like this…

- Page (“word”)
  - Language
    - **Etymology**
    - **Pronunciation**
    - Morphological category (verb, noun, male, female...)
      - Headword line
      - Definitions (senses)
        - Sense 1
        - …
      - Translations
        - Sense 1
        - ...
    - Morphological category
    - …
  - Language
    - …

Layout of a typical Wikidata item

On Wikidata, data about concepts is stored in items, consisting of:

- One (or no) label per language
- One (or no) description per language
- Any number of aliases per language
- One (or no) sitelink per sister project
- Multiple statements

Lexicographical data will be stored in a new entity type because they need a specific structure.

Future: the Lexeme on Wikidata

On Wikidata, a Lexeme, like item, will have its own page, with:

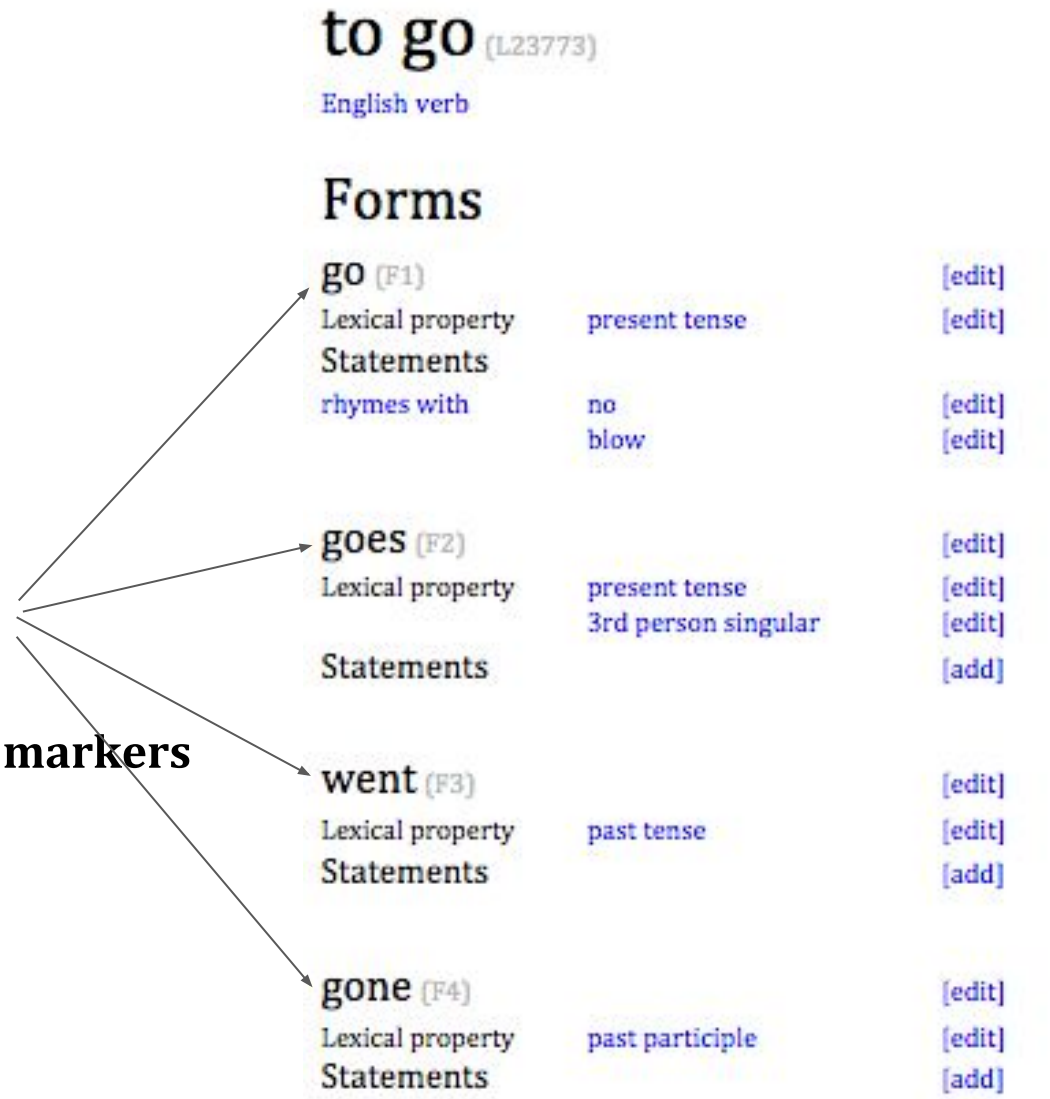
- 1 Lemma (mostly for display purposes, e.g. infinitive form)
- 1 Lexical category (e.g. verb, noun, etc., from Item space)
- 1 Language (e.g. English, German, etc., from Item space)
- Multiple Forms, each with
  - 1 Representation (the actual string)
  - Multiple Grammatical markers
  - Multiple Statements (e.g. region, period, pronunciation, etc.)
- Multiple Senses
  - 1 Gloss per language (=definition)
  - Multiple Statements (e.g. translations, synonyms, connotation, register, usage example, refers-to-concept)
- Multiple Statements (e.g. derived-from, pronunciation, region, period, etc.)

Lexeme ID

- **1 Lemma**
- **1 Lexical category**
- **1 Language**
- Multiple **Forms**, each with
  - 1 **Representation**
  - Multiple **Grammatical markers**
  - Multiple **Statements**
- Multiple **Senses**
  - **1 Gloss per language**
  - Multiple **Statements**
- Multiple **Statements**

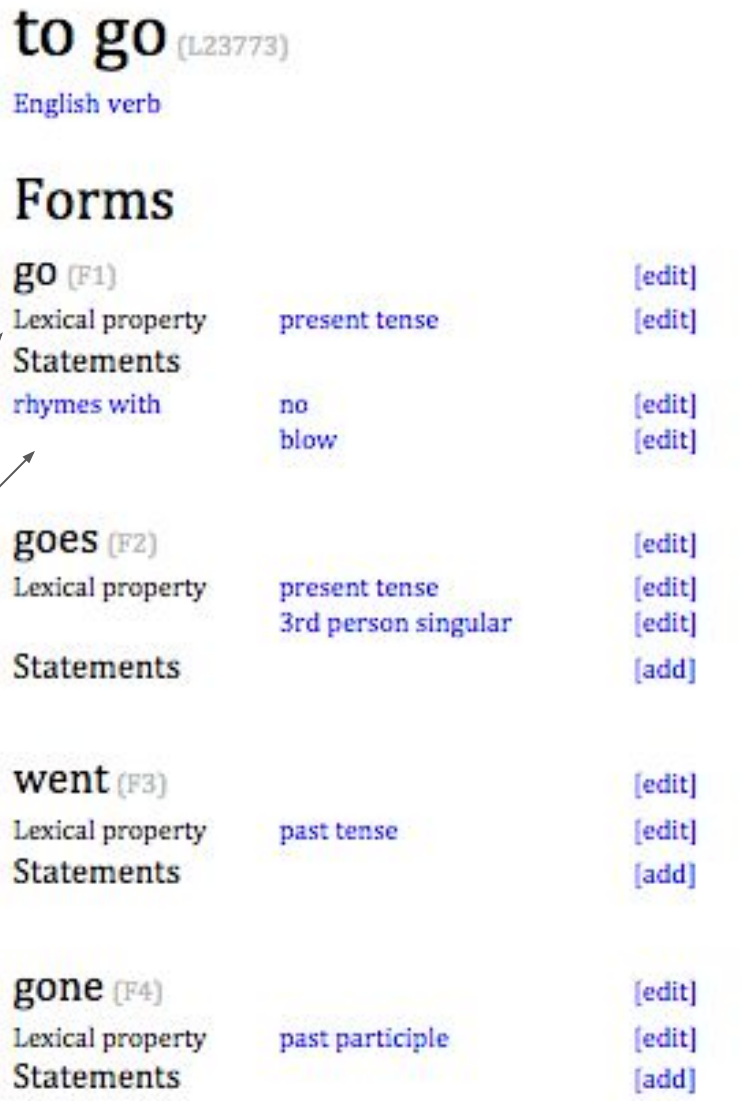
Lexeme

- **1 Lemma**
- **1 Lexical category**
- **1 Language**
- Multiple **Forms**, each with
  - 1 **Representation**
  - Multiple **Grammatical markers**
  - Multiple **Statements**
- Multiple **Senses**
  - **1 Gloss per language**
  - Multiple **Statements**
- Multiple **Statements**



Lexeme

- **1 Lemma**
- **1 Lexical category**
- **1 Language**
- Multiple **Forms**, each with
  - 1 **Representation**
  - Multiple **Grammatical markers**
  - Multiple **Statements**
- Multiple **Senses**
  - **1 Gloss per language**
  - Multiple **Statements**
- Multiple **Statements**

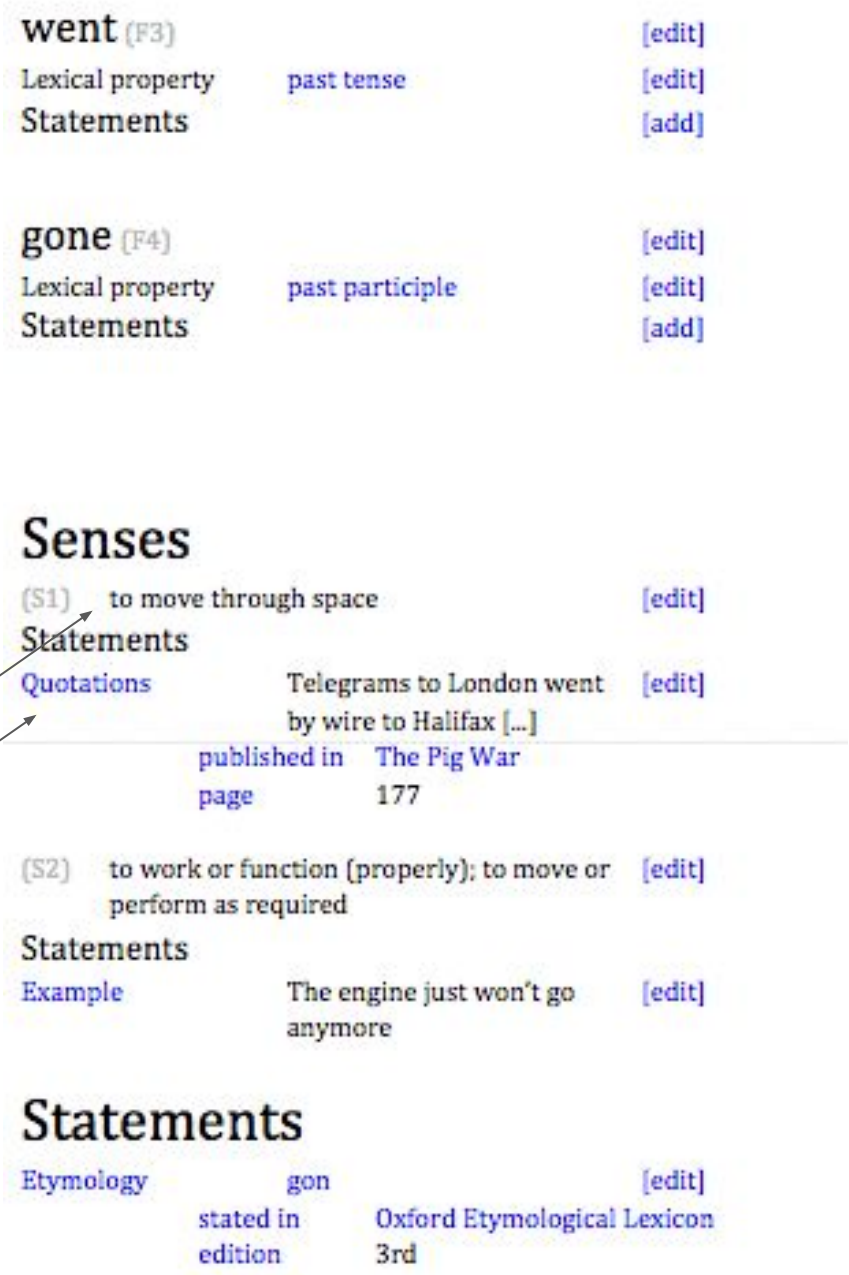


Lexeme

- **1 Lemma**
- **1 Lexical category**
- **1 Language**
- Multiple **Forms**, each with
  - 1 **Representation**
  - Multiple **Grammatical markers**
  - Multiple **Statements**
- Multiple **Senses**
  - **1 Gloss per language**
  - Multiple **Statements**
- Multiple **Statements**

Lexeme

- **1 Lemma**
- **1 Lexical category**
- **1 Language**
- Multiple **Forms**, each with
  - 1 **Representation**
  - Multiple **Grammatical markers**
  - Multiple **Statements**
- Multiple **Senses**
  - **1 Gloss per language**
  - Multiple **Statements**
- Multiple **Statements**



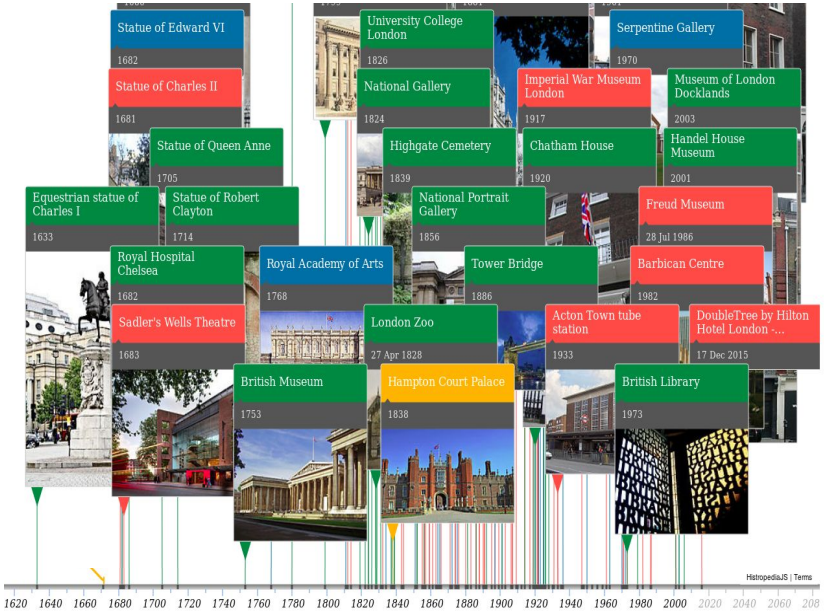
Lexeme

- **1 Lemma**
- **1 Lexical category**
- **1 Language**
- Multiple **Forms**, each with
  - 1 **Representation**
  - Multiple **Grammatical markers**
  - Multiple **Statements**
- Multiple **Senses**
  - **1 Gloss per language**
  - Multiple **Statements**
- Multiple **Statements**

Useful external tools that already use Wikidata

Reusing Wikidata data :

- Telescope infobox on English Wikipedia
- Histropedia, visual timelines
- Inventaire, keep an inventory of your books using Wikidata metadata
- Yle, Finnish broadcasting company using Wikidata for their tags



Editing :

- Wikidata Games, fun and easy ways to edit Wikidata
- WikiShootMe : find the items where pictures are missing (very useful for Wiki Loves Monuments for example)
- Mix'n'Match: connecting other databases to Wikidata

Our plan

Phase alpha
Automatic interwiki links on Wiktionary (connects pages with the same name on the different Wiktionaries to each other)

Phase beta
Create new entity types (for Lexeme, Form, Sense and Embedded)

Phase gamma
Provide arbitrary access from data on Wiktionary
(the ability to dynamically include any data from Wikidata on any Wiktionary page)

Phase delta
Improve the display of Wiktionary data on Wikidata
(create compact views, handle multiple representations)

See also the detailed tasks

How can you help?

- Check how your use cases fit into the data model
- Tell us more about your use cases
- Tell us more about specific needs in your language
- Create a Wikidata project on your local Wiktionary (example on French one) and add the link on the list
- Share any ideas you have about projects, tools, improvements that could be made using Wikidata!

Frequently asked questions

- Why will this project be useful for Wiktionary editors?
- How can you put lexical information into a database?
- Why do we need the data to be machine-readable?
- Will all the information be transferred from Wiktionary to Wikidata?
- Will we be forced to make use of Wikidata's data?
- Will it be more difficult to contribute to Wiktionary?

You will find the answers to these questions on our FAQ page.

Additional questions?

The talk page of the project waits for your comments and ideas!

You'll be able to follow our progress on the Wikidata Weekly Summary.

Thanks for your feedbacks :)

Wikidata development team