# Import of the *Etymological lexicon of modern Breton* by Victor Henry from Wikisource into Wikidata lexicographical data

•••

Envel Le Hir

www.lehir.net

ContribuLing
April 22, 2022

# Program

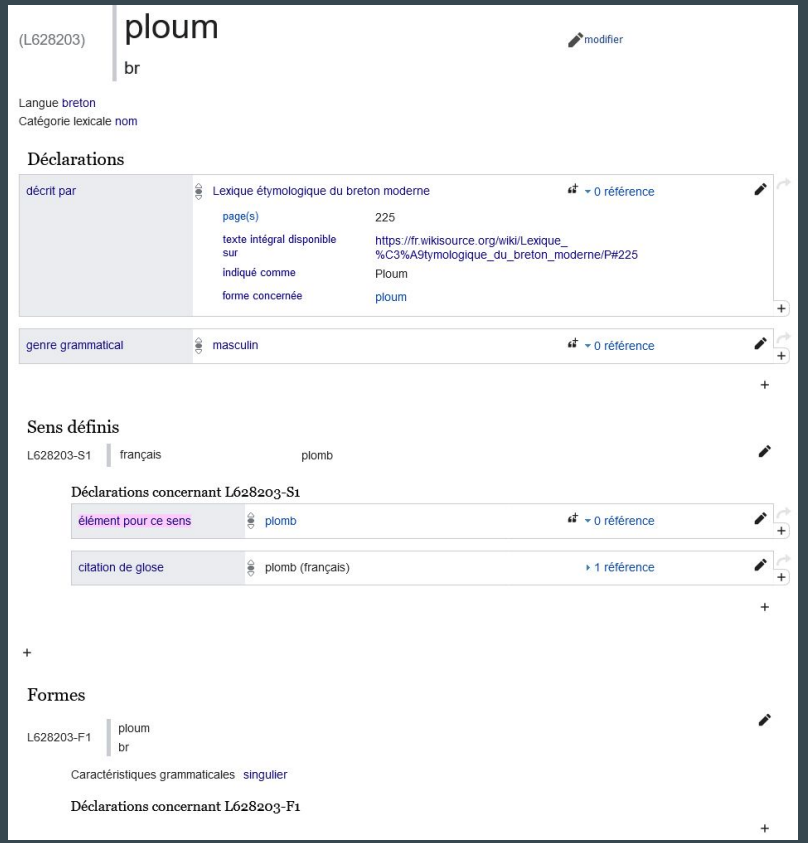
- Wikidata and lexicographical data
- Import project

# Wikidata and lexicographical data: history

- 2002 : launch of Wiktionary
- 2012 : launch of Wikidata, developed by Wikimedia Germany
- 2013 : first formal discussions to modelize lexicographical data in Wikidata
- 2018 : first version of lexicographical data deployed on Wikidata, halting of developments, maintenance
- 2019 : *Wikidata et Wiktionnaire : retour sur un échec annoncé* [*Wikidata and Wiktionary: overview of a foretold failure*], by Pamputt
- 2020 : announce of *Abstract Wikipedia* and *WikiFunctions*
- December 2021 : Basque and Bengali Wiktionaries are the first to have direct access to lexicographical data from Wikidata (T212843)

# Wikidata and lexicographical data: data model

- Same model for all languages
  - id L...
  - lemmas
  - language
  - lexical category
  - statements
    - etymology
    - relevant properties (example: grammatical gender)
    - references
  - senses
    - glosses
  - forms
    - inflections with grammatical features
- Example : [Lexeme:L628203]

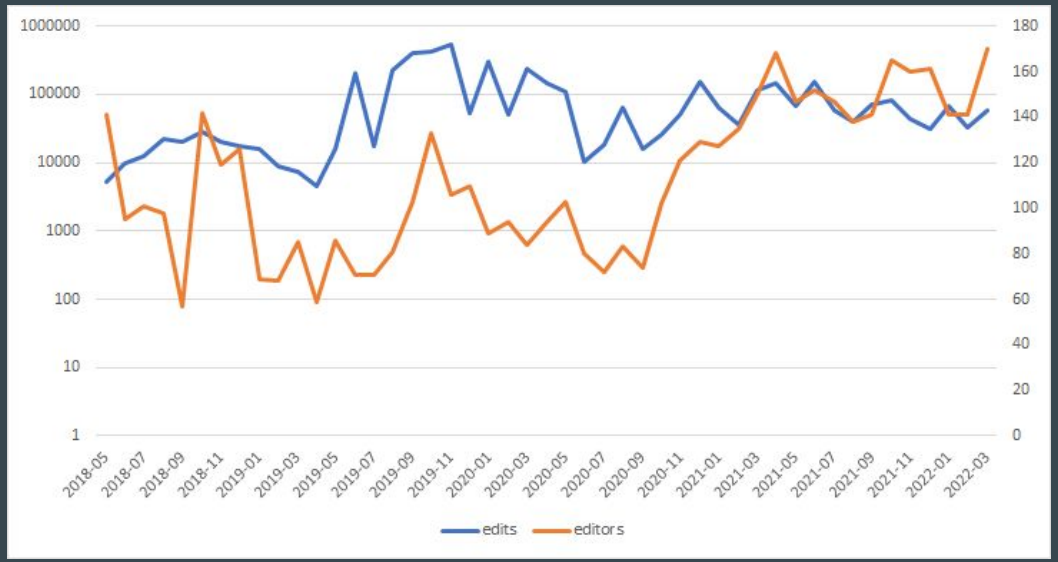


# Wikidata and lexicographical data: license

- Data on Wikidata is under CC0 license, equivalent to public domain.
- It is not possible to import data under more restrictive licenses into Wikidata. Example: Wiktionary under CC BY-SA 3.0 license.
- However, data from Wikidata can be reused without restriction.

# Wikidata and lexicographical data: community

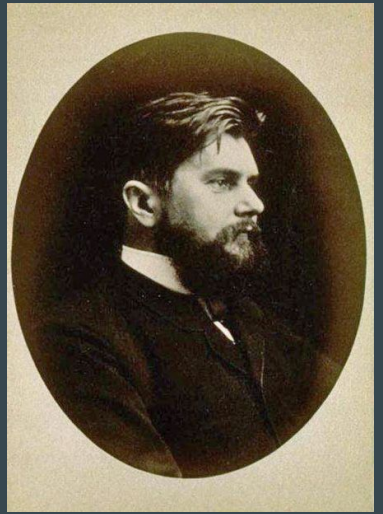
- Project on Wikidata: Wikidata:Lexicographical data
- Telegram group: https://t.me/joinchat/ICn09hkymb2dwpFKwGo5uA
- Ecosystem of tools operated by the community, including:
  - Ordia: statistics about lexemes on Wikidata
  - Wikidata Lexeme Forms: to create and update lexemes forms on Wikidata
  - Lexemes Challenge: collaborative weekly challenge to improve the coverage of lexemes on Wikidata




Source: quarry.wmcloud.org/query/63914

# Import project of the *Etymological lexicon of modern Breton*

- *Etymological lexicon of modern Breton*
  - dictionary, in French, about the Breton language
  - written by Victor Henry (1850-1907), French linguist
  - published in 1900, in the public domain
  - available on Wikisource
- Idea, in Summer 2021, with Nicolas Vigneron, to use it to improve Breton lexemes on Wikidata (less than 300 at that time).



Victor Henry

# Wikisource

- Wikimedia project of digital library, composed of works in the public domain or under free license.
- A work is available as scans and then, after OCR, in wikicode format for proofreading by volunteers.
- After proofreading, the work can be read and exported in many formats (HTML, PDF, EPUB, etc.).

# Wikisource

**Ploué**, s. m., campagne, village : autrefois, et dans les noms de lieux (*Plou-*), « paroisse, communauté d'habitants », corn. *plui* > *plu* > *plew*, cymr. *plwyf* > *plwy*, vbr. *pluio*. Empr. lat. *plēbēs*.
**Ploum**, s. m., plomb, corn. *plom*, cymr. *pliom*. Empr. lat. *plumbum*.
**Plouz**, s. m., fétu. Empr. fr. ancien *pelous* « velu ».
**Plû**, s. m., plume, mbr. *pluff* et *pluvenn*, corn. *pliv*, cymr. *pluf* > *plu*. Empr. lat. *plūma*.

```
'''Ploué''', s. m., campagne, village : autrefois, et dans les noms de lieux (Plou-), « paroisse, communauté d'habitants », {{abréviation|corn.|cornique}} plui>plu > plew, cymr. plwgf^> plicy, vbr. pluio. Empr. {{abréviation|lat.|latin}} plèbes.

'''Ploum''', s. m., plomb, {{abréviation|corn.|cornique}} plom, cymr. plie m. Empr. {{abréviation|lat.|latin}} plumbum.

'''Plouz''', s. m., fétu. Empr. fr. ancien pelous « velu ».

'''Plû''', s. m., plume, {{abréviation|mbr.|moyen-breton}} pluff et pluoenn, {{abréviation|corn.|cornique}} pliv, cymr. pluf> plu. Empr. {{abréviation|lat.|latin}} pluma.
```

**Ploué**, s. m., campagne, village : autrefois, et dans les noms de lieux (Plou-), « paroisse, communauté d'habitants », corn. plui>plu > plew, cymr. plwgf^> plicy, vbr. pluio. Empr. lat. plèbes.

**Ploum**, s. m., plomb, corn. plom, cymr. plie m. Empr. lat. plumbum.

**Plouz**, s. m., fétu. Empr. fr. ancien pelous « velu ».

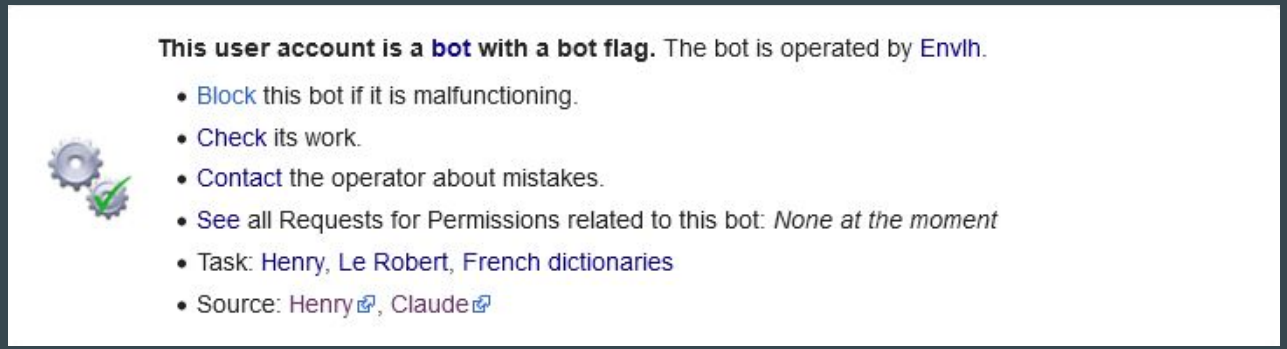
**Plû**, s. m., plume, mbr. pluff et pluoenn, corn. pliv, cymr. pluf> plu. Empr. lat. pluma.

# Transformation of wikicode into a format compatible with Wikidata

- Use of Mediawiki API to retrieve the content of the book from Wikisource
- Parsing of wikicode with a Python script, including:
  - normalization (example: quotes)
  - adjustment to each lexical category (example: for a name, the grammatical gender)
  - dialects
- Reports
  - list of lexemes
  - list of errors
  - letters frequencies (unigrams, bigrams)
- Iterative process
  - fixes in Wikisource
  - new parsing, with new reports generation

# Import: Wikidata bot

- Request for permission, required before running a bot on Wikidata
- Test phase : between 50 and 250 edits
- User page of the robot: User:EnvlhBot



# Import

- ~300 existing lexemes in Breton, ~3700 to import
- Some lexemes already exist in Wikidata, they must not be recreated.
- Use of the property *described by source* (P1343):

# Manual adjustments

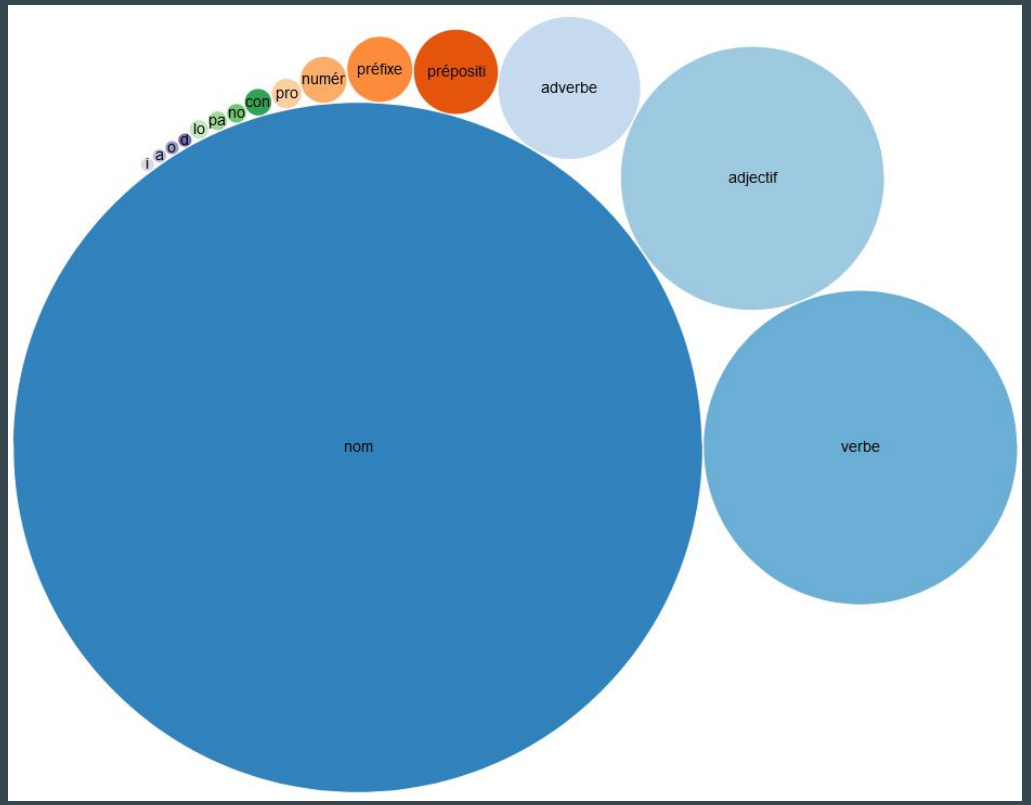
- Important step: an import is never perfect
- First workshop in January 2022:
  - 7 participants
  - introduction to lexicographical data on Wikidata and to the import project
  - collaborative work on lexemes, especially the addition of senses
- Some tasks:
  - missing etymologies: https://w.wiki/55gM
  - missing senses: https://w.wiki/55gN
  - adding other forms, using other dictionaries

# Documentation

- Lack of documentation for projects in the Wikimedia movement
- Documentation of the import project, in various complementary ways:
  - Source code (already reused by another project): https://github.com/envlh/henry
  - Blog post: https://www.lehir.net/how-we-imported-the-etymological-lexicon-of-modern-breton-from-wikisource-into-wikidata-lexicographical-data/
  - Workshops and conferences: January 2022, today, etc.
- Documentation on Wikidata:
  - Dedicated page for the Breton language: https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Documentation/Languages/br

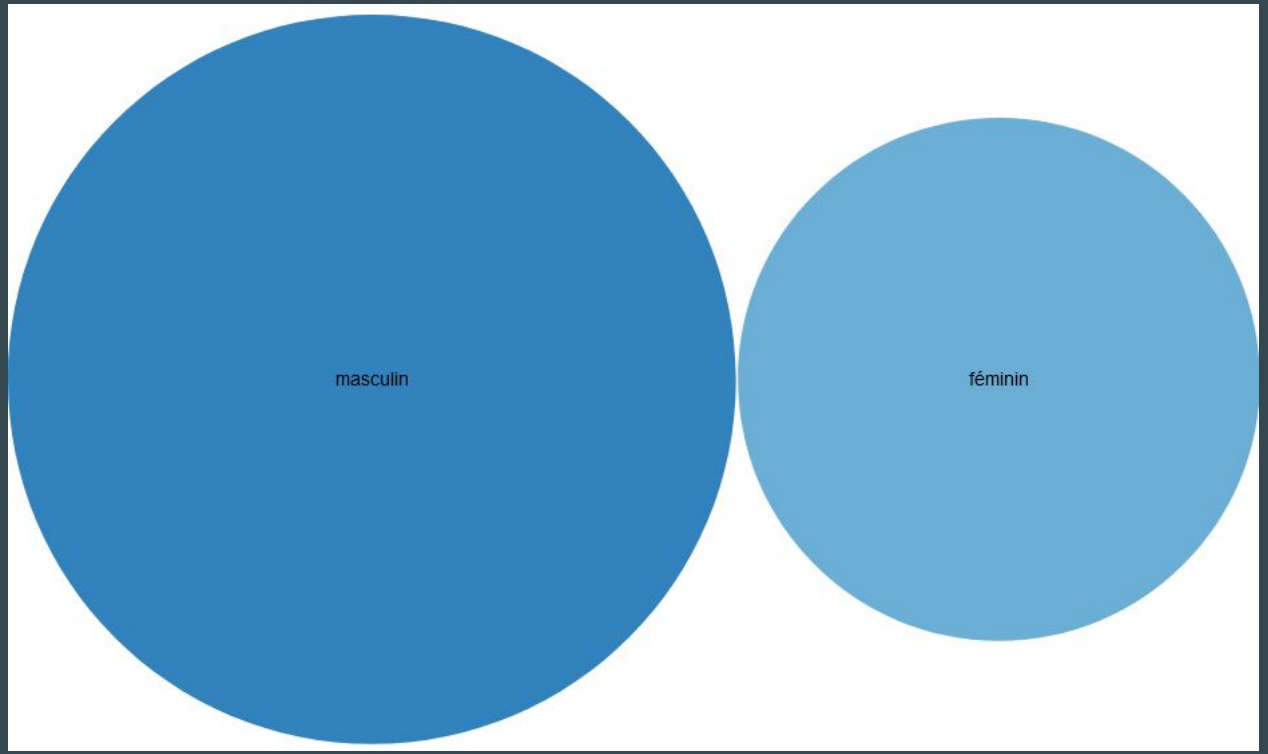
# Example: lexical categories

- https://w.wiki/55gb



# Example: grammatical genders of names

- https://w.wiki/55gd



# Example: links to concepts on Wikidata

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [de] allemand | Möbel<br>Möbelstück | Bett<br>Bettchen | Hocker | Stuhl<br>Stühlchen | Tisch<br>Tischchen | Schreibtisch | Bücherregal |
| [en] anglais | furniture | bed | stool | chair | table | desk | bookcase |
| [eu] basque | altzari | ohe | | | mahai | | |
| [bn] bengali | | | টুল | | টেবিল | | |
| [nb] bokmål | møbel | seng | krakk | stol | bord | skrivebord | bokhylle<br>bokreol |
| [br] breton | arrebeuri | gwele | skabell | kador | taol | burev | armel-levrioù |
| [hr] croate | | | | | stol | | |
| [da] danois | møbel | seng | taburet | stol | bord | skrivebord | bogreol |
| [es] espagnol | mobiliario | cama | taburete | silla | mesa | escritorio | librería |
| [eo] espéranto | meblo | lito | tabureto | segho / segxo / seĝo | tablo | librotablo | libroŝranko |
| [fi] finnois | | | | tuoli | pöytä | | |
| [fr] français | meuble | lit | tabouret | chaise | table | bureau | bibliothèque |

- Extract from Lexemes Challenge #33

# Outcome

- Import is done
  - From less than 300 to more than 4,000 lexemes for the Breton  language in Wikidata.
  - All created lexemes have references.
  - Editors save time: they no longer need to create these lexemes and several statements are already filled.
- Difficulties
  - Lack of examples to develop around and to call the Wikibase API about lexemes.
  - Not everything can be automatically imported: there are still manual tasks to be done.
- Highlights
  - The process is documented and can be replicated (other dictionaries in Breton, other languages).
  - Data is in the public domain and can be easily queried (API, SPARQL).
  - The book has been improved on Wikisource.

# Questions

# Credits

- Envel Le Hir (c) CC BY-SA 4.0
- Photo of Victor Henry by Antoine Meyer, public domain