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

Envel Le Hir
CC BY-SA 4.0
[www.lehir.net](http://www.lehir.net)

ContributeLing
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](https://example.com/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](https://quarry.wmcloud.org/query/63914)
- Telegram group: [https://t.me/joinchat/ICn09hkymb2dwpFKwGo5uA](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 lexeme forms on Wikidata
  - **Lexemes Challenge**: collaborative weekly challenge to improve the coverage of lexemes on Wikidata

Source: [quarry.wmcloud.org/query/63914](https://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).
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.).
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. plom. Empr. lat. plumbum.

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

Plû, s. m., plume, mbr. pluff et pluenn, 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](User:EnvlhBot)

---

This user account is a **bot with a bot flag**. The bot is operated by Envlh.

- **Block** this bot if it is malfunctioning.
- **Check** its work.
- **Contact** the operator about mistakes.
- **See** all Requests for Permissions related to this bot: *None at the moment*
- Task: **Henry, Le Robert, French dictionaries**
- **Source**: Henry, Claude
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**):

<table>
<thead>
<tr>
<th><strong>described by source</strong></th>
<th><strong>Lexique étymologique du breton moderne</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>page(s)</td>
<td>225</td>
</tr>
<tr>
<td>full work available at URL</td>
<td><a href="https://fr.wikisource.org/wiki/Lexique_%C3%A9tymologique_du_breton_moderne/P#225">https://fr.wikisource.org/wiki/Lexique_%C3%A9tymologique_du_breton_moderne/P#225</a></td>
</tr>
<tr>
<td>stated as</td>
<td>ploum</td>
</tr>
<tr>
<td>subject form</td>
<td>ploum</td>
</tr>
</tbody>
</table>
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](https://w.wiki/55gM)
  - Missing senses: [https://w.wiki/55gN](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
  - 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](https://w.wiki/55gb)
Example: grammatical genders of names

- https://w.wiki/55gd
Example: links to concepts on Wikidata

<table>
<thead>
<tr>
<th>Language</th>
<th>Möbelstück</th>
<th>Bett</th>
<th>Hocker</th>
<th>Stuhl</th>
<th>Tisch</th>
<th>Schreibtisch</th>
<th>Bücherregal</th>
</tr>
</thead>
<tbody>
<tr>
<td>[de] allemand</td>
<td>Möbelstück</td>
<td>Bett</td>
<td>Hocker</td>
<td>Stuhl</td>
<td>Tisch</td>
<td>Schreibtisch</td>
<td>Bücherregal</td>
</tr>
<tr>
<td>[en] anglais</td>
<td>furniture</td>
<td>bed</td>
<td>stool</td>
<td>chair</td>
<td>table</td>
<td>desk</td>
<td>bookcase</td>
</tr>
<tr>
<td>[eu] basque</td>
<td>altzari</td>
<td>ohe</td>
<td>chair</td>
<td>mahai</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[bn] bengali</td>
<td></td>
<td></td>
<td>chair</td>
<td>mahai</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[nb] bokmål</td>
<td>mabel</td>
<td>seng</td>
<td>kradik</td>
<td>stol</td>
<td>bord</td>
<td>skrivebord</td>
<td>bokhylia</td>
</tr>
<tr>
<td>[br] breton</td>
<td>arrebeuri</td>
<td>gwele</td>
<td>skabell</td>
<td>kador</td>
<td>taol</td>
<td>burev</td>
<td>armel-leuenioù</td>
</tr>
<tr>
<td>[hr] croate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>stol</td>
</tr>
<tr>
<td>[da] dansois</td>
<td>mabel</td>
<td>seng</td>
<td>taburet</td>
<td>stol</td>
<td>bord</td>
<td>skrivebord</td>
<td>bogreol</td>
</tr>
<tr>
<td>[es] espagnol</td>
<td>mobiliario</td>
<td>cama</td>
<td>taburete</td>
<td>silla</td>
<td>mesa</td>
<td>escritorio</td>
<td>librería</td>
</tr>
<tr>
<td>[eo] esperanto</td>
<td>meblo</td>
<td>lito</td>
<td>tabureto</td>
<td>segho / segxo / sego</td>
<td>tablo</td>
<td>llibrotablo</td>
<td>librostranko</td>
</tr>
<tr>
<td>[fi] finnois</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>tuoli</td>
<td>pöytä</td>
<td></td>
</tr>
<tr>
<td>[fr] français</td>
<td>meuble</td>
<td>lit</td>
<td>tabouret</td>
<td>chaise</td>
<td>table</td>
<td>bureau</td>
<td>bibliothèque</td>
</tr>
</tbody>
</table>

- 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