Import du *Lexique étymologique du breton moderne* de Victor Henry depuis Wikisource dans les données lexicographiques de Wikidata

•••

Envel Le Hir

www.lehir.net

ContribuLing
22 avril 2022

Programme

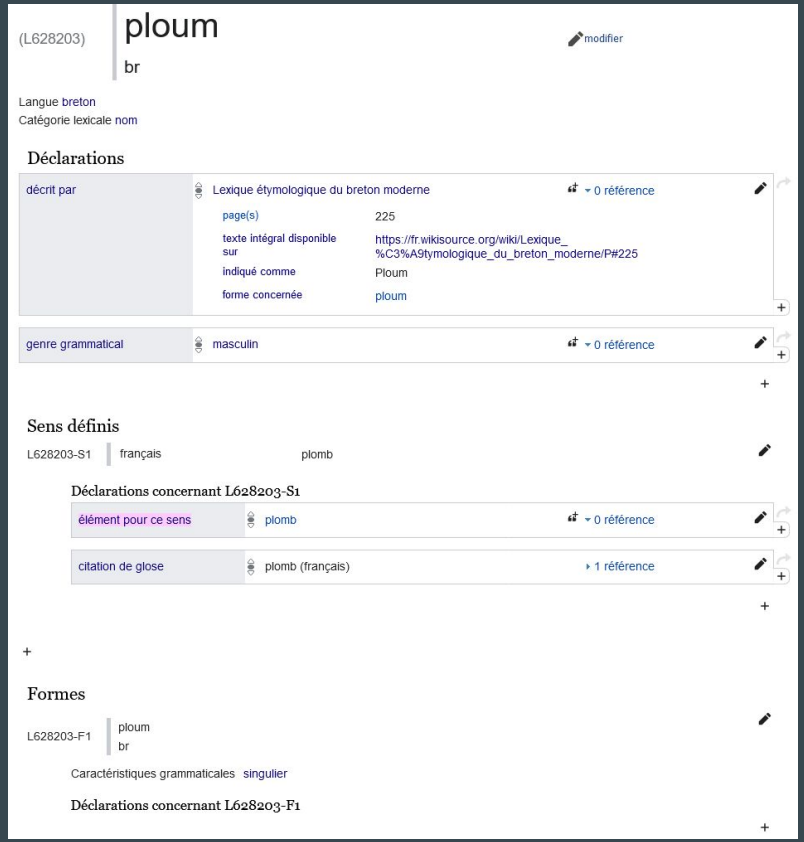
- Wikidata et les données lexicographiques
- Le projet d’import

Wikidata et les données lexicographiques : historique

- 2002 : naissance du Wiktionnaire
- 2012 : naissance de Wikidata, développé par Wikimédia Allemagne
- 2013 : premières discussions formelles sur l'intégration de données lexicographiques dans Wikidata
- 2018 : déploiement de la première version des données lexicographiques dans Wikidata, arrêt des développements, maintenance
- 2019 : *Wikidata et Wiktionnaire : retour sur un échec annoncé*, par Pamputt
- 2020 : annonce de *Abstract Wikipedia* et *WikiFunctions*
- décembre 2021 : les wiktionnaires basque et bengali sont les premiers à pouvoir accéder directement aux données lexicographiques de Wikidata (T212843)

Wikidata et les données lexicographiques : modèle de données

- Modèle commun pour toutes les langues
  - identifiant unique L...
  - lemmes
  - langue
  - catégorie lexicale
  - déclarations
    - étymologie
    - propriétés pertinentes (exemple : genre grammatical)
    - sources
  - sens
    - gloses
  - formes
    - flexions avec leurs caractéristiques grammaticales
- Exemple : [Lexeme:L628203]

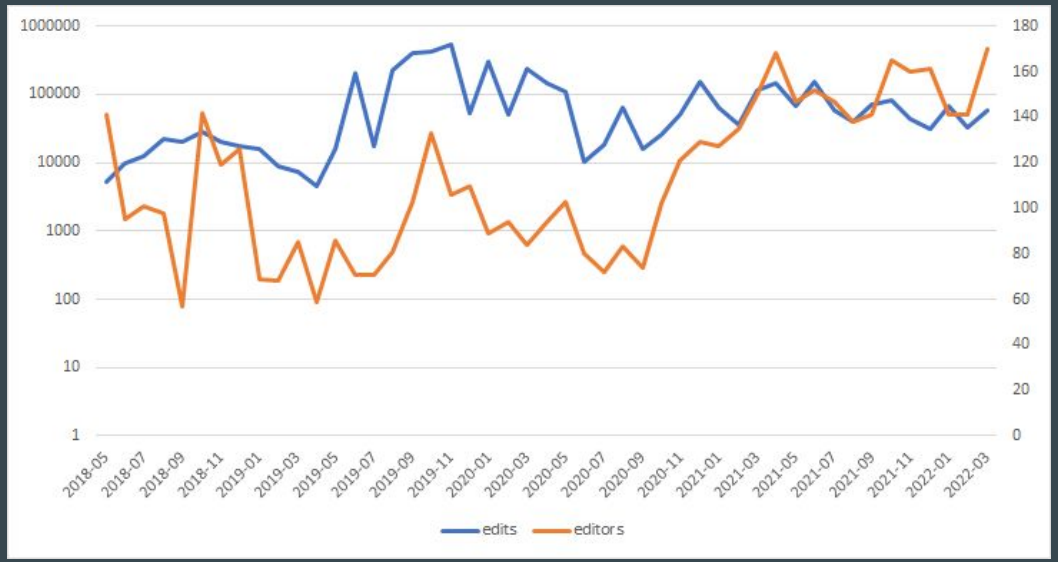


Wikidata et les données lexicographiques : licence

- Toutes les données de Wikidata sont sous licence CC0, équivalent du domaine public.
- Il n’est pas possible d’importer dans Wikidata des données qui sont sous des licences plus restrictives. Exemple : Wiktionnaire sous licence CC BY-SA 3.0.
- En revanche, les données de Wikidata peuvent être réutilisées sans restriction.

Wikidata et les données lexicographiques : communauté

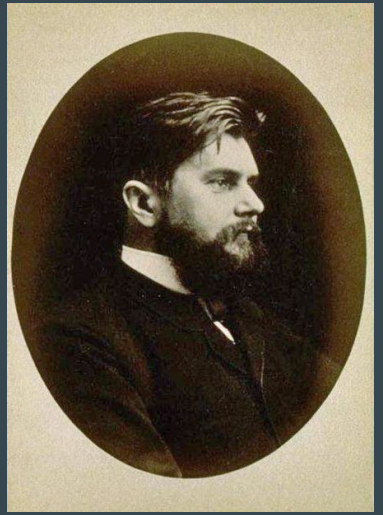
- Projet sur Wikidata : Wikidata:Lexicographical data
- Groupe Telegram : https://t.me/joinchat/ICn09hkymb2dwpFKwGo5uA
- Écosystème d'outils maintenus par la communauté, dont :
  - Ordia : statistiques sur les lexèmes dans Wikidata
  - Wikidata Lexeme Forms : pour créer et mettre à jour les formes des lexèmes dans Wikidata
  - Lexemes Challenge : défi collaboratif hebdomadaire pour améliorer la couverture des lexèmes dans Wikidata



Source : quarry.wmcloud.org/query/63914

Projet d'import du *Lexique étymologique du breton moderne*

- *Lexique étymologique du breton moderne*
  - dictionnaire, en français, à propos de la langue bretonne
  - écrit par Victor Henry (1850-1907), linguiste français
  - publié en 1900, dans le domaine public
  - disponible sur Wikisource
- Idée, à l'été 2021, avec Nicolas Vigneron, de l'utiliser pour améliorer les lexèmes en breton dans Wikidata (moins de 300 à ce moment-là).



Victor Henry

Wikisource

- Projet Wikimédia de bibliothèque numérique, composée d'œuvres dans le domaine public ou sous licence libre.
- Un ouvrage est disponible sous forme de scans puis, après une étape d'OCR, au format wikicode pour transcription et relecture par les bénévoles du projet.
- Après relecture, l'ouvrage peut être consulté et exporté dans de nombreux 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 du wikicode en un format compatible avec Wikidata

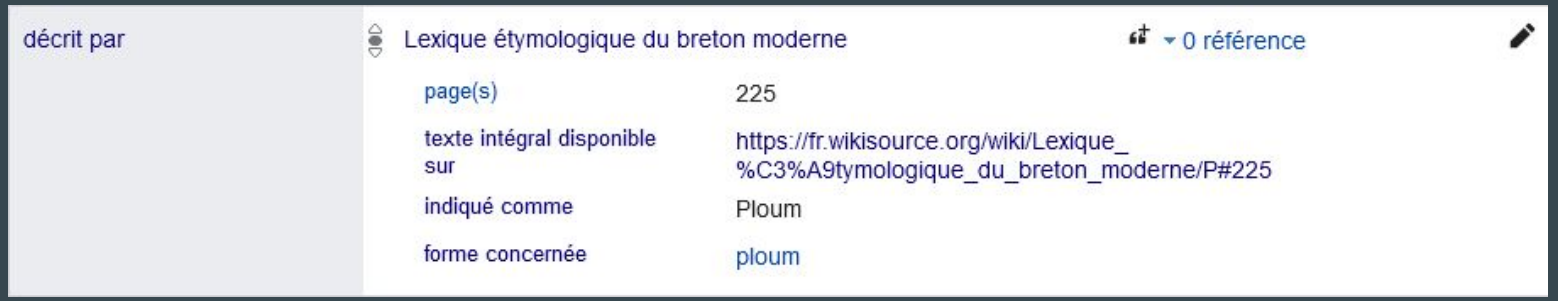
- Utilisation de l'API Mediawiki pour récupérer le contenu de Wikisource
- Parsing du wikicode avec un script Python, dont :
  - normalisation (exemple : apostrophes)
  - adaptation à chaque catégorie lexicale (exemple : pour un substantif, le genre grammatical)
  - dialectes
- Rapports
  - liste des lexèmes
  - liste d'erreurs
  - fréquences des lettres (unigrammes, bigrammes)
- Processus itératif
  - corrections dans Wikisource
  - nouveau parsing, avec génération des rapports

Import : bot Wikidata

- Demande de permission, indispensable avant d'utiliser un robot sur Wikidata
- Phase de test de l'import : entre 50 et 250 modifications
- Page d'utilisateur du bot : User:EnvlhBot

Import

- ~300 lexèmes existants en breton, ~3700 à importer
- Certains lexèmes existent déjà dans Wikidata, il ne faut pas les recréer.
- Utilisation de la propriété *décrit par* (P1343) :



Ajustements manuels

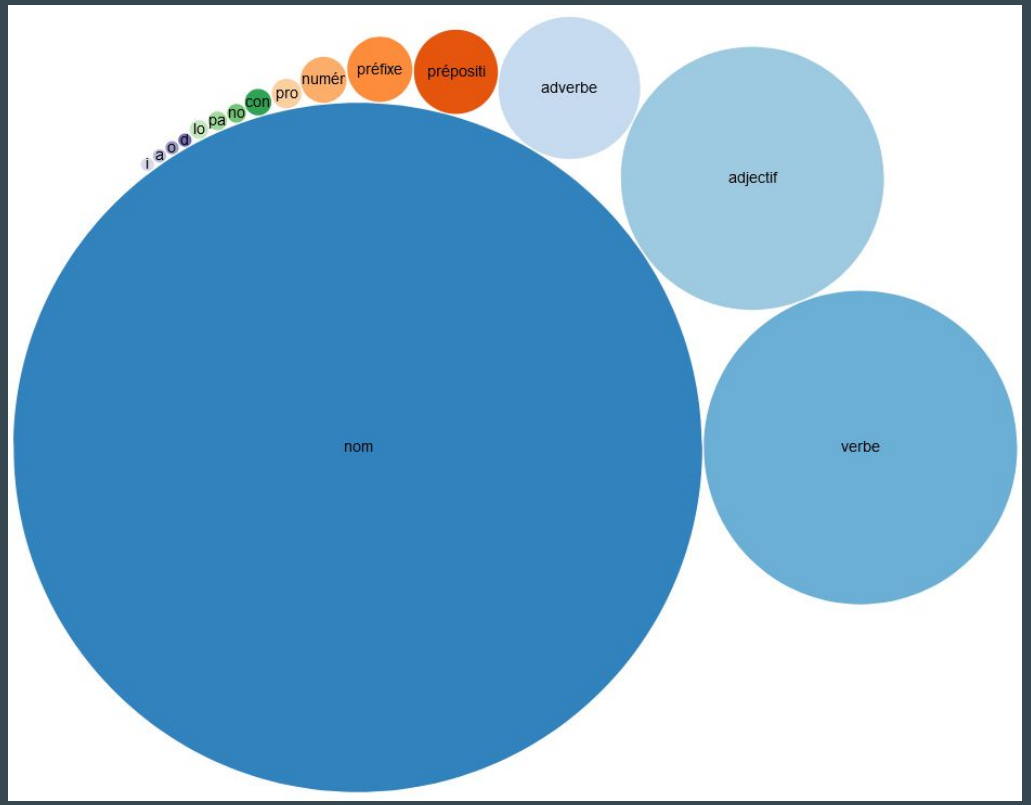
- Étape à ne pas négliger : un import n'est jamais parfait
- Premier atelier en janvier 2022 :
  - 7 participants
  - présentation des données lexicographiques dans Wikidata et du projet d'import
  - travail collaboratif sur les lexèmes, en particulier l'ajout de sens
- Quelques tâches :
  - étymologies manquantes : https://w.wiki/55gM
  - sens manquants : https://w.wiki/55gN
  - ajouter d'autres formes, à l'aide d'autres dictionnaires

Documentation

- Manque de documentation des projets dans le mouvement Wikimédia
- Documentation du projet d'import, sous plusieurs formes complémentaires :
  - Code source (déjà réutilisé par un autre projet pour le norvégien) : https://github.com/envlh/henry
  - Billet de blog : https://www.lehir.net/how-we-imported-the-etymological-lexicon-of-modern-breton-from-wikisource-into-wikidata-lexicographical-data/
  - Ateliers et conférences : janvier 2022, aujourd'hui, etc.
- Documentation de Wikidata :
  - Page dédiée pour le breton : https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Documentation/Languages/br

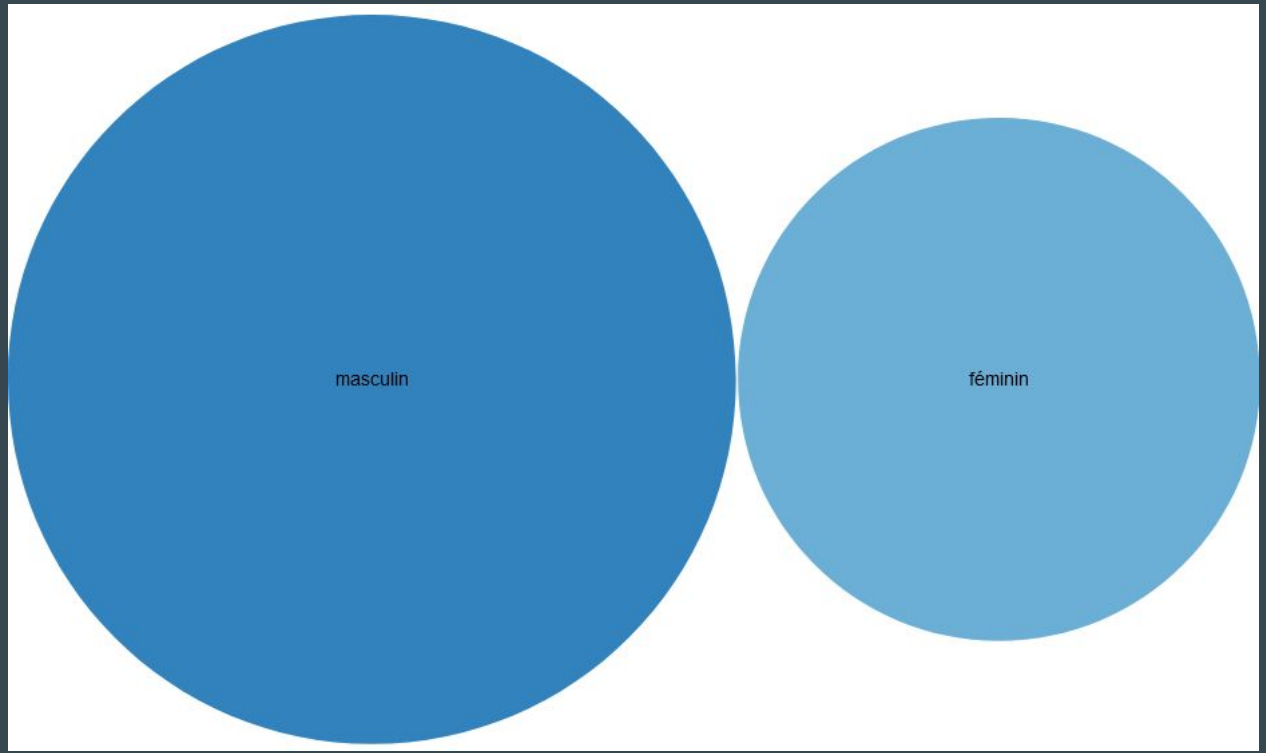
Exemple : catégories lexicales

- https://w.wiki/55gb



Exemple : genres grammaticaux des noms

- https://w.wiki/55gd



Exemple : liens avec les concepts de Wikidata

[de] allemand	Möbel Möbelstück	Bett Bettchen	Hocker	Stuhl Stühlchen	Tisch 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 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

- Extrait du Lexemes Challenge #33

Bilan

- Import réalisé
  - On passe de moins de 300 à plus de 4000 lexèmes en breton dans Wikidata.
  - Les lexèmes créés sont tous sourcés.
  - Les contributeurices gagnent du temps : il n'est plus nécessaire de créer ces lexèmes et de nombreuses données sont déjà saisies.
- Difficultées rencontrées
  - Manque d'exemples pour le développement du code et les appels à l'API Wikibase sur les lexèmes.
  - Impossible de tout importer automatiquement : il reste du travail manuel.
- Points forts
  - Le processus est documenté et peut être répliqué (autres dictionnaires en breton, autres langues).
  - Les données sont dans le domaine public et peuvent être facilement interrogées (API, SPARQL).
  - L'ouvrage a été amélioré sur Wikisource.

Questions

Crédits

- Envel Le Hir (c) CC BY-SA 4.0
- Photo de Victor Henry par Antoine Meyer, domaine public