Wikidata et Wiktionnaire : l’échec a bien eu lieu !

Pamputt, 20 ans du Wiktionnaire, 23-24 mars 2024

Plan

Rapide historique

Avis légal de la Fondation sur les données lexicographiques

Licence CC0 pour les données lexicographiques ?

Wikimedia Commons

5 ans après, où en est-on ?

Rapide historique

- 22 mars 2004 : création du Wiktionnaire en français
- 30 octobre 2012 : création de Wikidata
- 22 février 2018 : il est demandé à la communauté Wikidata d'approuver la licence CC0 pour les données lexicographiques.
- 23 mai 2018 : les données lexicographiques sont activées sur Wikidata (seuls les lexèmes et les formes sont disponibles)
- 18 octobre 2018 : les sens sont activés

Avant l'arrivée des données lexicographiques

Wikimania 2016 : rencontre entre des contributeurs au Wiktionnaire et des développeurs Wikidata

« Il y avait de leur part (dév. Wikidata) une véritable motivation à **faire évoluer techniquement** les Wiktionnaires et à enfin faire avancer la discussion avec les communautés de contributeurs, qui n'ont pas été particulièrement attirés par la question jusqu'ici. »

Création de wikt:Projet:Coopération/Wikidata

Avant l'arrivée des données lexicographiques

De nombreuses propositions de modèles de données

Date de début	Auteur(s) principal(aux)
2013-02	JAn Dudík (cswikt), This, that and the other (enwikt), Darkdadaah (frwikt)
2013-06	Denny (WMDE)
2013-07	Micru (ca, wikidata), Francis Tyers (iswikt)
2013-08	Denny (WMDE)
2013-09	Ivadon (dewikt)
2013-10	Bigbossfarin (enwikt)
2014-10	GPHemsley (enwiki, wikidata)
2015-05	Denny (WMDE)

Toutes les propositions ont été traduites en français (sauf 2013-09, en allemand) pour pouvoir en débattre plus facilement (~63 ko de texte)

Avis légal de la Fondation Wikimedia (17/02/2018)

m:Wikilegal/Lexicographical_Data/fr

Informations lexicographiques **non protégées** par le droit d'auteur :

- Lemmes (le mot lui-même)
- Définitions de jargon ou de mots avec des expressions figées
- Prononciations
- Informations grammaticales (masculin/féminin, nom/verbe/…, …)
- Collocations et expressions figées (exemple : « lourd silence », « pluie battante » ne sont pas protégeables)

Avis légal de la Fondation Wikimedia (17/02/2018)

m:Wikilegal/Lexicographical_Data/fr

Informations lexicographiques **protégées** par le droit d'auteur :

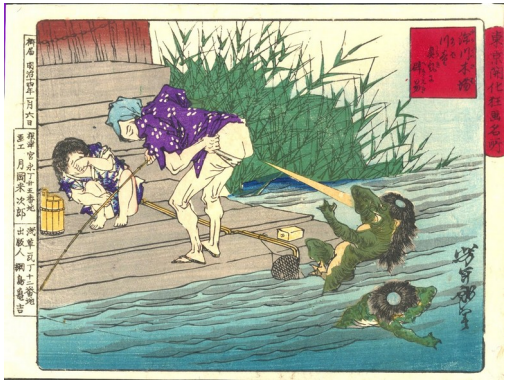
- Microstructure et macrostructure (l'organisation du dictionnaire).
  Est-ce qu'importer toutes les pages de catégories du Wiktionnaire (élément de la macrostructure) en tant qu'élément dans Wikidata est autorisé ?
- Définitions
- Étymologies
- Informations pragmatiques (façon dont le mot est utilisée, s'il est archaïque, etc)
- Informations encyclopédiques
- Exemples de phrases (dans le cas d'exemples originaux)

Avis légal de la Fondation Wikimedia (17/02/2018)

m:Talk:Wikilegal/Lexicographical Data

Questions ouvertes. Quid de ?

- Choix d'une illustration associée à une définition
- Liste de tous les mots d'un dictionnaire
- Thésaurus



Un prout

Licence CC0 pour les données lexicographiques (22/02/2018)

Wikidata:Project_chat#Adding_the_Lexeme_namespace_to_the_licensing_footer_text

Discussions vives entre les soutiens de la licence CC0 permettant une réutilisation la plus large possible et des contributeurs/contributrices aux Wiktionnaires (principalement) qui voient Wikidata comme un projet concurrent qui va se faire en parallèle des Wiktionnaires

La discussion apparaît comme un vote alors qu'il est très probable que les développeurs de Wikidata aient prévus d'appliquer la licence CC0 quel que soit le résultat

Licence CC0 pour les données lexicographiques (22/02/2018)

Wikidata:Project_chat#Adding_the_Lexeme_namespace_to_the_licensing_footer_text

Je considère que vous faites un fork
du Wiktionnaire dans Wikidata
avec votre propre calendrier.
Noé

Licence CC0 pour les d…

Wikidata:Project_chat#Adding_th…

Je considère que vous fai…
du Wiktionnaire dans Wi…
avec votre propre calendrier.
Noé

Ce projet a commencé sous le nom de « Données structurées pour le Wiktionnaire ». Quelque part en chemin, le but a changé, et il est devenu « Wikidata pour les données lexicographiques ». Je pense que ce changement était une erreur. Je ne pense pas qu'il puisse y avoir deux projets lexicographiques rivaux chez Wikimedia.
Jheald

Licence CC0 pour les d… …ues

Wikidata:Project_chat#Adding_th…

Ce projet a commencé sous le nom de « Données structurées pour le Wiktionnaire ». Quelque part en chemin, le but a changé, et il est devenu « Wikidata pour les données lexicographiques ». …it une erreur.

Je considère que vou… …projets du Wiktionnai… avec vot…

C'est exactement ce que nous craignions depuis le début sur le Wiktionnaire : que Wikidata commence à traiter les données lexicographiques sans même se soucier de consulter les personnes qui créent et gèrent déjà des données lexicographiques sur Wikimedia chaque jour. Compte tenu de la situation des licences et du manque flagrant de communication, deux projets parallèles vont travailler sur les mêmes problèmes, mais séparément.

Metaknowledge

Licence CC0 pour les données lexicographiques (22/02/2018)

Au bout d'une semaine de discussions, strakhov a fait les comptes

Pour la licence CC0 : 26 personnes (~5910 modifications sur les Wiktionnaires, dont 5055 par un seul contributeur)

Contre : 9 personnes (~158 027 modifications sur les Wiktionnaires)

Abstention : 3 personnes (~127 000 modifications sur les Wiktionnaires)

Licence CC0 [illegible] lexicographiques

Au [illegible]

[illegible]

Co [illegible]

Abstentio [illegible]ktionnaires)

Je voulais juste dire que le seul contributeur significatif dans Wiktionary soutenant cette proposition est apparemment VIGNERON. Je ne vois pas en quoi l'antagonisme d'un projet entier est bon pour nous, même s'il attire d'autres personnes de l'extérieur. Je ne m'oppose pas à la « structuration » de Wiktionary, car je trouve le travail effectué là aussi assez inefficace, mais j'essaierais d'impliquer ces communautés au lieu de les contourner ici par la force brute. Leur demander et leur donner ce dont ils ont besoin. Si une partie importante des communautés Wiktionary pense que leur travail est plagié ou utilise une mauvaise licence par cette approche des lexèmes de Wikidata et de la licence CC0, il est peut-être temps de repenser la proposition

strakhov

Intermède : comment ça se passe avec Commons ?

c:Commons:Structured_data

2017-2019 : projet « Données structurées » sur Wikimedia Commons

Stockage des (méta)données relatives aux fichiers multimédias de manière structurée

Le *backend* de Commons est migré sur Wikibase

Intermède : comment ça se passe avec Commons ?

c:Commons:Structured_data

2017-2019 : projet « Données structurées » sur Wikimedia Commons

Stockage des (méta)données relatives aux fichiers multimédias de manière structurée

Le *backend* de Commons est migré sur Wikibase

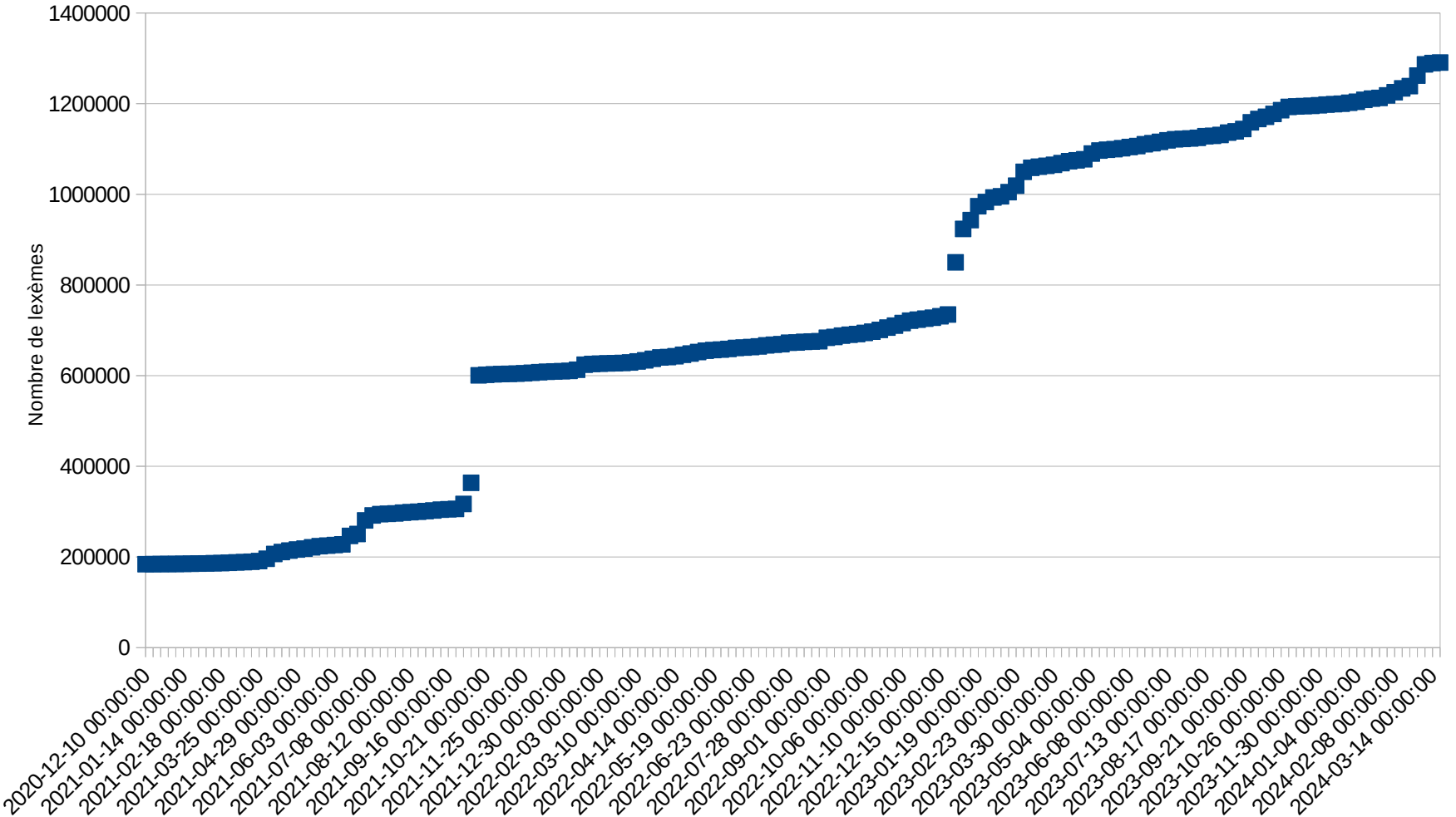
Collaboration entre les développeurs de Wikidata et la communauté de Commons
Pourquoi pas aussi avec le Wiktionnaire ?

Finalement, les lexèmes arrivent

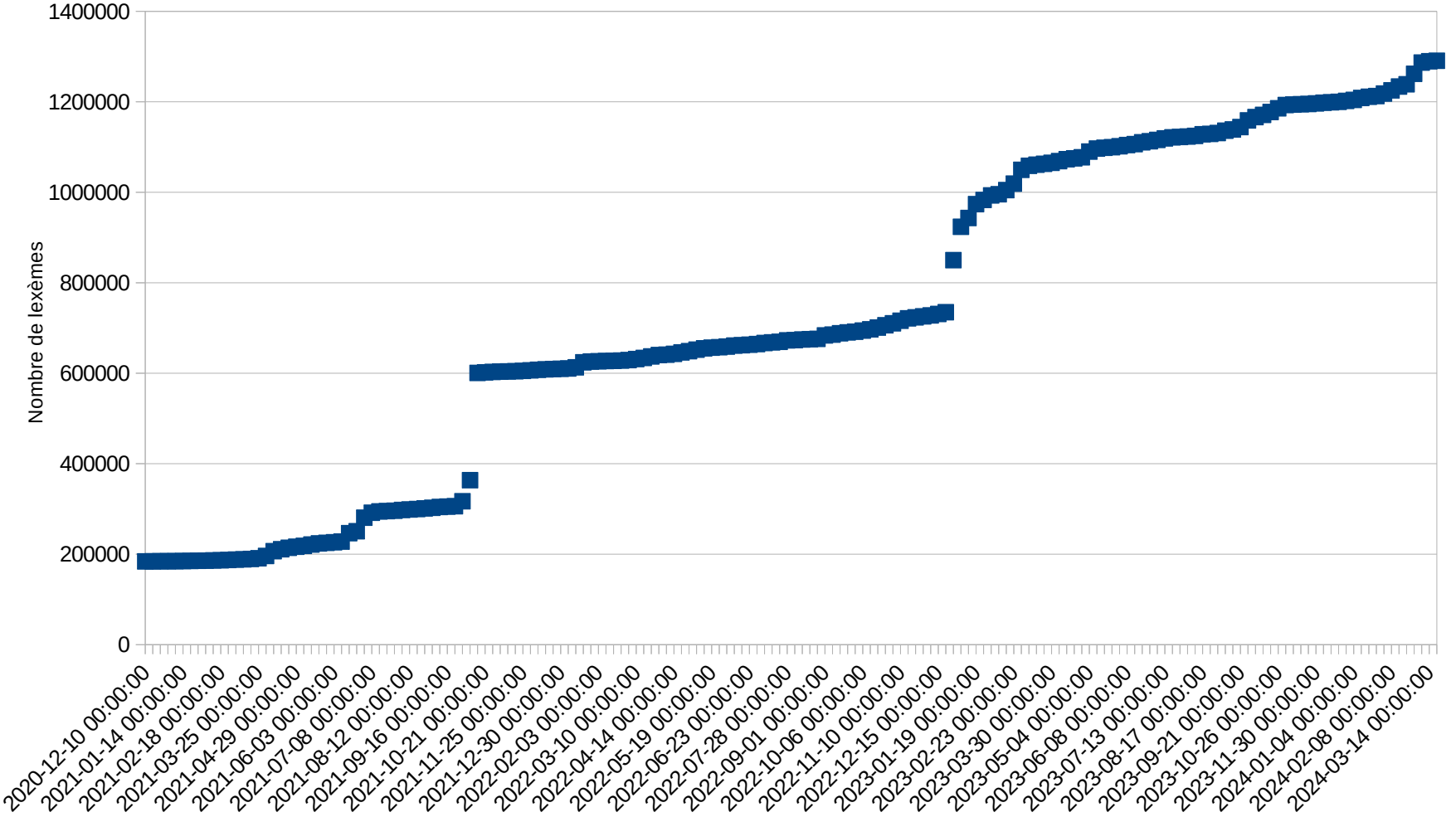
22 mai 2018 : activation des lexèmes et des formes
il est demandé de ne pas importer de données des Wiktionnaires

22 septembre : l'import de données par bot est autorisé par l'équipe Wikidata

18 octobre : les sens sont activés



Mais, ~72 % des lexèmes n’ont pas de sens… (en mars 2024)



https://grafana.wikimedia.org/d/UHV96YJGk/wikidata-datamodel-lexemes?orgId=1

Des personnes des Wiktionnaires ?

6 ans après le lancement des lexèmes :

- **156** utilisateurs ont fait **plus de 1000 contributions**

- **22** ont fait **plus de 100 000 contributions** (dont 8 bots)

- Parmi les **20 plus gros contributeurs** : **3** ont contribué significativement (>1000 edits) à un Wiktionnaire (Jon Harald Søby (en, no), VIGNERON (fr, br), Vis M (en, ml))

https://quarry.wmcloud.org/query/44611

Respect des licences ?

Crainte initiale au sujet de l'import des données du Wiktionnaire (CC by-sa 3.0/4.0) dans Wikidata (CC-0)

- Septembre 2019 : importation par bot de lexèmes russes. Les sens, synonymes, etc ne sont pas importées. Une discussion est initiée sur ruwikt pour relicensier ruwikt en CC0. La discussion n'aboutit pas.
- Septembre 2019 : l'association basque officialise la demande bot pour importer des sens à partir d'un dictionnaire en basque sous licence propriétaire (le propriétaire du bot est aussi l'éditeur du dictionnaire)

https://quarry.wmcloud.org/query/44611

Utilisation des données lexicographiques

13 décembre 2021 : test des données lexicographiques sur bnwikt et euwikt)
21 juin 2022 : accès aux données lexicographiques sur tous les Wiktionnaires (T309 593)

Mars 2024 : à ma connaissance, seul euwikt utilise les données lexicographiques (exemple)
Après l'intérêt du début (2017-2018) des gros Wiktionnaires pour les données lexicographiques, il semble que l'intérêt a disparu depuis.

Pour quelles raisons ?

Utilisation des données lexicographiques

- Difficulté technique : besoin de connaissance en Lua

- Accès aux données lexicographiques non intuitifs
  il faut saisir l'identifiant du lexème, de la forme ou du sens pour accéder aux données ; ça rend le wikicode cryptique
  ex : {{conjugaison|L47}} pour obtenir un tableau de conjugaison du verbe « aimer » (L47)

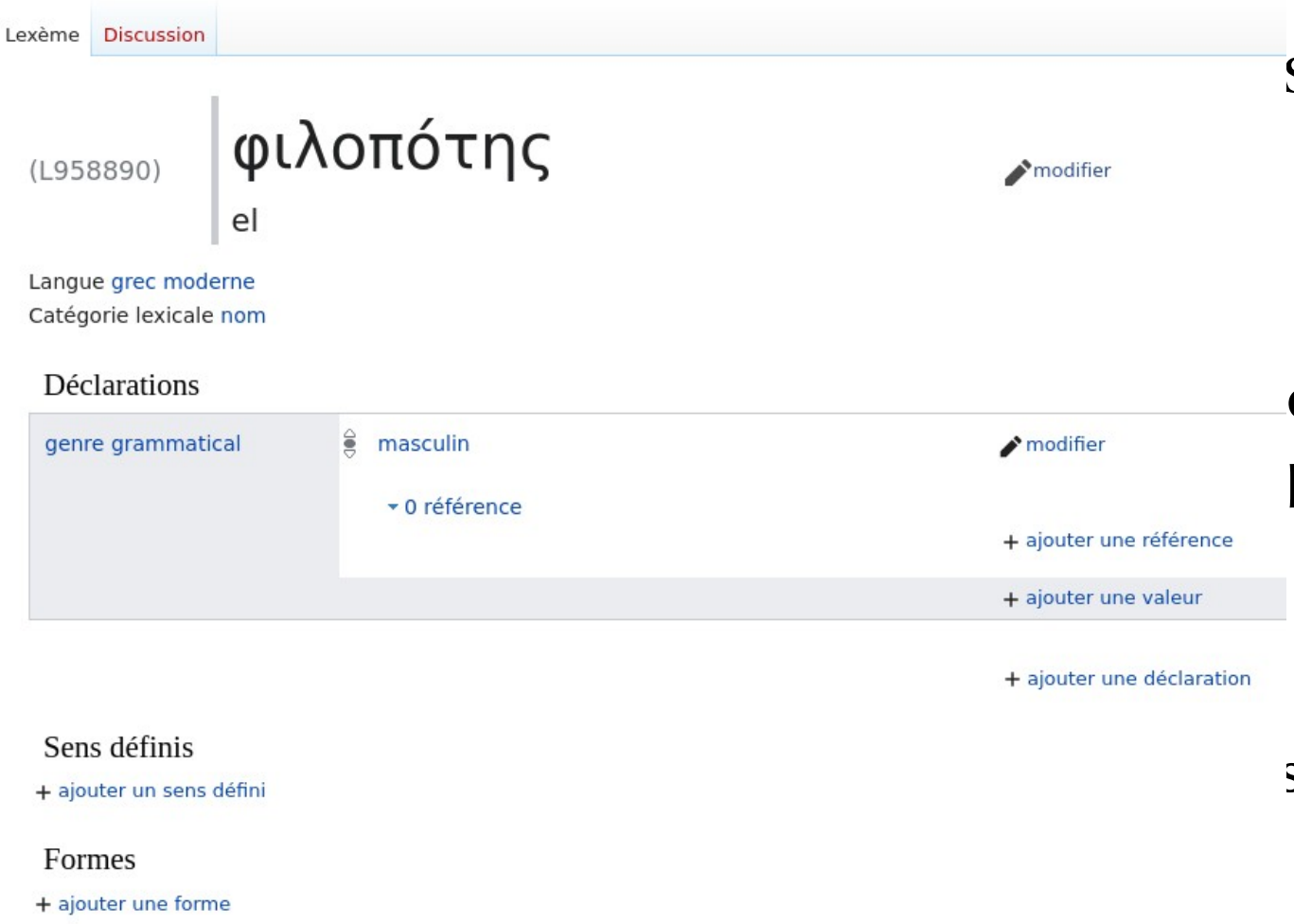
- Données lexicographiques trop incomplètes et inégales

U s

- Difficu

- Accès
  il fa du sens
  pou ptique
  ex
  con

- Donné s



Conclusion

Wikidata est (toujours ?) vu comme un projet concurrent aux Wiktionnaires

Wikidata ne profite pas de l'expérience du contributorat des Wiktionnaires

Contributions isolées sur les données lexicographiques – pas encore de communauté structurée sur Wikidata (4 sujets ouvert depuis le début de l'année)

Mon avis en 2019 : Wikidata sera rempli de données purement grammaticale (déclinaison, conjugaison, …) qui pourront être réutilisées par les Wiktionnaire. Je ne crois pas au développement des sens.

En 2024 : les sens n'ont pas décollé en 5 ans (probablement du fait de la licence). Les Wiktionnaires ne réutilisent peu/pas les données lexicographiques. Les développeurs Wikidata ne communiquent plus sur les données lexicographiques.
Et « Cognate » n'est pas en forme (T326 432) et « I miss you » est cassé …

https://commons.wikimedia.org/wiki/File:Black-headed_weaver_(Ploceus_cucullatus_bohndorffi)_male_nest_building.jpg, CC by-sa 4.0

Merci pour votre attention

- Questions ?
- Commentaires ?
- Discussion ?

Wikidata : un Omegawiki bis ?

Omegawiki avait le but de fournir un dépôt central de données lexicographiques. Ce projet est objectivement un échec.

- Trop technique : utile pour ajouter/réutiliser des informations purement techniques (déclinaisons, conjugaison, ...)

- Attention aux discussions. Une seule langue (anglais) (ou presque) conduit à une vision anglo-centrée et exclut *de facto* des contributeurs. Comment assurer et encoder la diversité linguistique ?