Wikidata talk:Lexicographical data

Overview	Documentation	Development	Tools	Support Wiktionary	for Hov	w to help	Lexemes	Disc
Wikidata:Lexicograp	ohical data							
						-	org/w/index.php?title= ographical+data%2FHe	-
Lexico	graphical	data	_				ader%2Ftext))
Place used to di	scuss any and all aspe blicy and proposals, inc	l items.	Start a new discussion (https://www.wikidata.org/w/index.p hp?title=Wikidata_talk:Lexicographical_data&action=edit§ion =new)					

On this page, old discussions are archived. An overview of all archives can be found at this page's **archive index**. The current archive is located at 2018/10.

Project	Lexicographical	Administrators'	Development	Translators'
chat	data	noticeboard	team	noticeboard
Request	Requests	Requests	Bot	Requests
a query	for deletions	for comment	requests	for permissions
Property	Properties	Partnerships	Interwiki	Bureaucrats'
proposal	for deletion	and imports	conflicts	noticeboard

Contents

Regional words Over 10,000 in 1 language (French)! L:L21070 should not exist Pronunciation respelling for English Q2428747 or Q1084 Notice for tool developers A few lexemes disappeared **Query Lexemes in the Query Service** Senses are now part of Lexicographical Data

Regional words

Do we have a property that would work for marking where regional words/senses are used? For example, wikt:en:car hire says Australia, New Zealand and Britain. wikt:en:wisht says Cornwall and Devon.

The closest I can find is indigenous to (P2341) but that sounds really weird to me. There's used by (P1535) and valid in place (P3005) too but they don't seem appropriate either. Maybe we need a lexeme-specific property?

- Nikki (talk) 15:28, 27 September 2018 (UTC)

Use "Language Australian English (Q44679), New Zealand English (Q44661), British English (Q7979)" in lieu of "Language English (Q1860)" for wikt:en:car hire; and create Wikidata items for "English in Cornwall" (or Anglo-Cornish (Q3299229)? see this (https://tools.wmflabs.org/wikidata-todo/tree.html?lang= en&q=Q15839081&rp=279)) and "English in Devon" for wikt:en:wisht. Visite fortuitement prolongée (talk) 15:30, 29 September 2018 (UTC), 15:42, 29 September 2018 (UTC), 15:45, 29 September 2018 (UTC), 15:54, 29 September 2018 (UTC)

I think that is a really bad idea and I'm strongly opposed to it. They are English words found in English dictionaries, the language should be English. It's not realistic to expect people to create and maintain numerous identical lexemes just because a sense is not used by everyone. It would be a nightmare when a word has lots of senses (e.g. wikt:en:pot, most of those senses would belong on the regional lexemes too) or when a word is used in a lot of regions (e.g. wikt:en:pants, which would need at least 11 lexemes for the first sense if every region has its own). A very similar issue is that some senses are specific to certain subjects (e.g. sports, nautical terms) and we need a way to mark those. Not everyone uses those senses either, but that doesn't mean they should be separate lexemes. - Nikki (talk) 16:15, 30 September 2018 (UTC)

I'm very much with Visite fortuitement prolongée here. If they are regional dialect *words*, they have to be marked as such. The argument about words with multiple senses is a misunderstanding and highly misleading: these can be marked as regional in the exact same way on the relevant sense (duh!) when senses become accessible in mid-october. Now, I'm not saying a "dialect" subproperty wouldn't be potentially useful, but as long as one doesn't exist and items for the dialects do, "language"+dialect is a perfectly reasonable property combination to use. Circeus (talk) 18:09, 2 October 2018 (UTC)

Nobody said they shouldn't be marked. My argument is that treating dialects as independent languages and therefore requiring independent lexemes is an awful way to do it. I am strongly in favour of statements on senses instead. We can't mark individual senses on a lexeme without using a property. The only language codes senses have are the ones which mark

the language that the gloss (the text describing the meaning) is written in. For example, the German word "blau" could have an en-gb gloss which uses "colour" and an en-us gloss which uses "color". - Nikki (talk) 21:14, 2 October 2018 (UTC)

Since nobody has made any suggestions for which property to use, I've proposed a new one: Wikidata: Property proposal/location of sense usage. -Nikki (talk) 12:01, 6 October 2018 (UTC)

Over 10,000 in 1 language (French)!

According to Wikidata: Lexicographical data/Statistics there are 10221 in fr; all other languages are still under 10,000. The total is getting close to 30,000. ArthurPSmith (talk) 18:13, 4 October 2018 (UTC)

I think there are two languages: English with an entity for every 1-5 letter word! Congrats Arthur! --- Jura 18:48, 4 October 2018 (UTC)

Actually for me English is about twice bigger than French because every lexeme in English contains complete set of forms while I can hardly find lexeme with complete set of forms in French. Describing English and others languages usually goes "deep", while French goes "flat". For example average size of last 10 lexemes created in English is 855,1 while for last 10 French lexemes it is 449,6. For Polish the same number is 17332,2. KaMan (talk) 07:29, 5 October 2018 (UTC)

It's hard to say. The good thing about the Polish ones is that the generally have statements about the entities (e.g. gender) which others lack. For other languages, some are missing because the relevant properties haven't been created yet. I'm not sure if we are quite ready for forms yet. I think the bug report about the automatic creation of F1 is still open and it's layout is probably only going to happen next year.

An interesting find from the approach for English is that lookup for short lemmas (and forms) starts getting saturated. --- Jura 07:42, 5 October 2018 (UTC)

L:L21070 should not exist

Hello, what do you think about L:L21070? To me, it should not exist since International Phonetic Alphabet (Q21204) is not a language. If we want to use analogy, it is closer to a script than a language. Any opinion? Pamputt (talk) 18:42, 4 October 2018 (UTC)

I'm not sure if it has much use as a unique entity for Q21204. For French, please see Wikidata talk:Lexicographical data#Count number of vowels above. --- Jura 18:46, 4 October 2018 (UTC)

Well it was created in French, then changed to IPA if you look at the history. I agree IPA is not a language. ArthurPSmith (talk) 18:48, 4 October 2018 (UTC)

It's not a French word either. Or I would say it's as much a French word as it is a Spanish one. Pamputt (talk) 18:55, 4 October 2018 (UTC)

Well, the lexical category is correct for French. Wikidata talk:Lexicographical data#Count number of vowels explains why it's needed there. For entities for characters, there is a discussion at Wikidata: Requests for permissions/Bot/GZWDer (flood) 3. --- Jura 18:59, 4 October 2018 (UTC)

I strongly disagree that /o/ is a French word, this is a IPA symbol. Otherwise, what about [o], \o\, ...? What would be the meaning of this word (we can continue in French if you want)? Just because you need this lexeme for storing some information does not mean it is a French word. Pamputt (talk) 19:05, 4 October 2018 (UTC)

We can continue the discussion in the thread above. As for the entity for IPA, I currently don't see the need. --- Jura 19:07, 4 October 2018 (UTC)

I have no specific opinion about the previous discussion. I just say that how it is doing now is not the good way. Pamputt (talk) 19:12, 4 October 2018 (UTC)

It's hard to say. Maybe Infovarious eventually details the scheme they have in mind. I don't want to be the person who doesn't contribute anything to his attempt and just fills the forum with idle comments. --- Jura 19:16, 4 October 2018 (UTC)

[UNDENT] Wasn't there a debate not too long ago about whether phonemes could be created as lexemes? I though the consensus was clearly against it? Circeus (talk) 22:35, 4 October 2018 (UTC)

Yes, @Circeus, Pamputt: See, Wikidata_talk:Lexicographical_data/Archive/2018/09#Is_phoneme_a_lexeme? KaMan (talk) 06:15, 5 October 2018 (UTC)

Looks more like a general discussion about the nature of phonemes. It seems to have ended with an unanswered question about how to store IPA. ---- Jura 06:39, 5 October 2018 (UTC)

From my POV it ended with consensus that phonemes should not be stored in lexeme namespace. And I see the same from this thread above. KaMan (talk) 06:44, 5 October 2018 (UTC)

I think we all agree that phonemes and letters aren't words, but that doesn't really help us building a structured database. Maybe you have a constructive suggestion for the point raised below? --- Jura 06:50, 5 October 2018 (UTC)

Store phoneme as Q-item. These are not lexemes. See close-mid back rounded vowel (Q862579) for example. Pamputt (talk) 09:43, 5 October 2018 (UTC)

How would the information be included? What would be the advantages over the approached favored by active contributors? How does it compare when they create entities for these? What would you suggest to them? Active contributors need to make editorial choices that store the information in an optimal way given the features of various entities. For L21070, it seems to be that the current lexical category is sub-optimal if not wrong. If this is coded as IPA, it should

probably be defined as a letter in that alphabet. --- Jura 08:35, 6 October 2018 (UTC)

Maybe those active contributors could explain to other active contributors why they need to add not-lexeme content into lexeme namespace? What can't be achieved without doing so? --Lexicolover (talk) 12:39, 6 October 2018 (UTC)

Pronunciation respelling for English

Some people like IPA to express spelling, but this doesn't seem to work for everyone. The result is that we have plenty of pronunciation files. Also some dictionaries attempt to express pronunciation in regular English language graphemes. See w: Pronunciation respelling for English. We could obviously attempt to store them in several formats for every word as it seems to be done for some Slavic languages, but a better solution might be to find a structured way to map the sounds to regular graphemes. How could this be done for English in a structured way with Wikidata? From your experience with creating entities such as Q- and L- ones, how would you go about it? --- Jura 06:39, 5 October 2018 (UTC)

Q2428747 or Q1084

Which items should be used for nouns? On L:L189, both are used. Pamputt (talk) 18:53, 4 October 2018 (UTC)

• I'd proceed as outlined above: Wikidata talk:Lexicographical data#Noun: one or several lexemes. I'm not really comfortable with the use made of P31. Either we should drop it or add a statement to every entity. --- Jura 18:56, 4 October 2018 (UTC)

I remeber it was stated recently in some threads that if lexical category is not set to proper noun (Q147276) then there is silent assumption that lexeme is common noun (Q2428747). Oh I found it. It was stated here https://www.wikidata.org/w/index.php?title=Wikidata_talk:Lexicographical_data&diff=prev&oldid=746423772 KaMan (talk) 07:37, 5 October 2018 (UTC)

Notice for tool developers

I've just documented the API which the Wikidata Lexeme Forms tool uses to search for potential duplicates of a lexeme you're about to create, over at User: Lucas Werkmeister/Wikidata Lexeme Forms #Duplicates. You're welcome to use the same API, either in your own tools that create lexemes (also to prevent creating duplicates), or as a stricter version of lexeme search (wbsearchentities with type=lexeme) which only returns exact matches on the lemma and language code. --Lucas Werkmeister (talk) 13:52, 5 October 2018 (UTC)

Does that mean we can finally use tools to create entities? --- Jura 08:36, 6 October 2018 (UTC)

@Jura1: I don't think that was really disallowed, as far as I understand... in the release announcement, Lea Lacroix (WMDE) asked us to refrain from any mass imports, which I understood to be more about bots - a tool that lets editors create individual lexemes, one by one, is okay as far as I understand. And she also suggested that we "wait a bit before building tools or scripts", but, well, I'm willing to deal with the risk of API changes breaking my tool :)

This API to find potential duplicates could also be used for a bot that automatically creates a bunch of lexemes, that's true, and I hope no one will build that for now – but I think there's still potential for some more tools like Wikidata Lexeme Forms, where editors create individual lexemes at a reasonable pace, and hopefully this API can help with that. --Lucas Werkmeister (talk) 08:18, 7 October 2018 (UTC)

Building tools on the top of the API is possible, just be aware that the system is not completely stable yet, some things may change, and require the tool developers to rewrite their code later. Lea Lacroix (WMDE) (talk) 08:39, 9 October 2018 (UTC)

A few lexemes disappeared

@Lea Lacroix (WMDE): I noticed strange thing. Lexemes from L20540 to L20543 just disappeared. They were neither merged, nor deleted. They for sure existed because they were Tracked in Phabricator created by me and I document every Polish lexeme I created on this page. As far as I remeber they were created 13 September (after datacenter switch) and they disappeared yesterday/today (after datacenter switch back). Could it be that this is related to https://wikitech.wikimedia.org/wiki/Switch_Datacenter#Schedule_for_2018_switch ? What should I do next with this case? KaMan (talk) 08:10, 11 October 2018 (UTC)

That's indeed very weird. Can you create a Phab ticket? I'll make sure that it's looked at soon. Lea Lacroix (WMDE) (talk) 08:24, 11 October 2018 (UTC)

@Lea Lacroix (WMDE): Ok, https://phabricator.wikimedia.org/T206743 KaMan (talk) 08:39, 11 October 2018 (UTC)

Maybe the same with Q56604439 (used at Q56604445#P734). --- Jura 09:09, 11 October 2018 (UTC)

Latest status: the data reappeared. Can you check again and see if you find other problems or missing items or Lexemes? Lea Lacroix (WMDE) (talk) 15:42, 11 October 2018 (UTC)

@Lea Lacroix (WMDE): I don't see problems now. Thanks to all involved, it was really fast response as I observed on phabricator. KaMan (talk) 16:00, 11 October 2018 (UTC)

Thanks for reporting this issue :) Lea Lacroix (WMDE) (talk) 16:41, 11 October 2018 (UTC)

Looks good. Thanks for fixing it. Seems like lexemes helped the global WMF project ensure stability! Q56604437 doesn't exist, but it might never had. Q56604439 was available on query server even when it had disappeared here. --- Jura 17:57, 11 October 2018 (UTC)

Query Lexemes in the Query Service

Hello all,

I'm very happy to announce that another important feature for Lexicographical Data has been deployed: the ability to **query Lexemes in the Query Service**.

Here are a few examples:

List of the longest words in English (https://query.wikidata.org/#SELECT%20DISTINCT%20%3FI%20%3Fword%20%3Flen%20WHERE%20%7B%0A%20

0000

Task T206743

Graph of Lexemes derived from L2087

%7B%0A%20%20%20%3Fl%20a%20ontolex%3ALexicalEntry%20%3B%20dct%3Alanguage%20wd%3AQ1860%20%3B%20wikibase%3Alemma%20%3Fword%0A%20%20%20BIND%28strle n%28%3Fword%29%20as%20%3Flen%29%20%20%0A%20%20%7D%20UNION%20%7B%0A%20%20%3Fl%20a%20ontolex%3ALexicalEntry%20%3B%20dct%3Alanguage%20wd%3A Q1860%20%3B%20ontolex%3AlexicalForm%2Fontolex%3Arepresentation%20%3Fword%0A%20%20BIND%28strlen%28%3Fword%29%20as%20%3Flen%29%20%20%0A%20%20%7D

%0A%7D%20%0Aorder%20by%20DESC%28%3Flen%29%20%0ALIMIT%2020), or in German (https://query.wikidata.org/#SELECT%20DISTINCT%20%3Fl%20%3Fword%20%3Flen%20WHER E%20%7B%0A%20%7B%0A%20%20%20%3Fl%20a%20ontolex%3ALexicalEntry%20%3B%20dct%3Alanguage%20wd%3AQ188%20%3B%20wikibase%3Alemma%20%3Fword%0A%20%20 %20BIND%28strlen%28%3Fword%29%20as%20%3Flen%29%20%20%0A%20%7D%20UNION%20%7B%0A%20%20%3Fl%20a%20ontolex%3ALexicalEntry%20%3B%20dct%3Alan guage%20wd%3AQ188%20%3B%20ontolex%3AlexicalForm%2Fontolex%3Arepresentation%20%3Fword%0A%20%20%20BIND%28strlen%28%3Fword%29%20as%20%3Flen%29%20%20% 0A%20%20%7D%0A%7D%20%0Aorder%20by%20DESC%28%3Flen%29%20%0ALIMIT%2020)

- Graph of all Lexemes (https://query.wikidata.org/#%23defaultView%3AGraph%0ASELECT%20%3Flexeme%20%3FlexemeLabel%20%3Ftarget%20%3 20%20%3Flexeme%20wdt%3AP5191%20%3Ftarget%3B%20wikibase%3Alemma%20%3FlexemeLabel.%0A%20%20%3Ftarget%20wdt%3AP5191*%20wd%3AL2087%3B%20wikibase%3Ale mma%20%3FtargetLabel.%0A%7D) derived from *wódr (L2087)
- Grammatical genders that are most used in lexicographical data in Wikidata (https://query.wikidata.org/#%23%20most%20common%20grammatical%20genders%0ASELECT%20%3Fgende r%20%3FgenderLabel%20%3Fcount%20WITH%20%7B%0A%20%20SELECT%20%3Fgender%20%28COUNT%28%3Flexeme%29%20AS%20%3Fcount%29%20WHERE%20%7B%0A%20%20 GROUP%20BY%20%3Fgender%0A%7D%20AS%20%25results%20WHERE%20%7B%0A%20%20INCLUDE%20%25results.%0A%20%20SERVICE%20wikibase%3Alabel%20%7B%20bd%3Aser viceParam%20wikibase%3Alanguage%20%22%5BAUTO LANGUAGE%5D%2Cen%22.%20%7D%0A%7D%0AORDER%20BY%20DESC%28%3Fcount%29)

The queries are based on the RDF mapping that you can find here. Feel free to help improving the documentation, so people can understand how to build queries out of Lexemes.

Thank you very much to Tpt who's been doing a huge part of the work by mapping Lexemes in RDF, and Smalyshev (WMF) who made the RDF dumps available and integrated in the Query Service.

Feel free to play with it, bring some of these ideas of queries to life, and let us know if you find any issue or bug. These can be stored as subtasks of this one on Phabricator. If you have questions, you can also ping Stas onwiki or on IRC.

Cheers, Lea Lacroix (WMDE) (talk) 08:06, 16 October 2018 (UTC)

Many thanks to all involved. That's great news and a lot of testing to do :) KaMan (talk) 08:33, 16 October 2018 (UTC)

- Cool, my first two queries ever (https://www.wikidata.org/w/index.php?title=Wikidata%3ALexicographical data%2FIdeas of queries&type=revision&diff=765519550&oldid=76065210 1) and they works :) KaMan (talk) 12:37, 16 October 2018 (UTC)
- Good work. BTW, there seems to be a licensing incompatibility with some of the schemes referenced in the triples. Can they be replaced with "wikibase:". Makes writing queries easier too. --- Jura 10:21, 16 October 2018 (UTC)
 - @Jura1: Could you explain a bit more about licensing? Smalyshev (WMF) (talk) 17:55, 16 October 2018 (UTC)
 - WMDE wanted lexemes to be CC0. If you are adding a primary mapping for key features of them to a scheme that isn't, somehow they fail that objective. Supposedly, you could still add it as a secondary mapping. It might also limit re-use of the software outside WikiMedia. --- Jura 13:15, 17 October 2018 (UTC)
 - @Jura1, Smalyshev (WMF): Are you talking about ontolex? The file at https://www.w3.org/ns/lemon/ontolex states a licence (dc:rights) value of CC-Zero, so it's fine. ArthurPSmith (talk) 15:17, 17 October 2018 (UTC)
 - That seems to be contradicted by statements elsewhere. Using the standard wikibase: seems preferable. --- Jura 15:24, 17 October 2018 (UTC)
 - (citation needed). Using a common standard makes federated guerying easier and so is preferable to using custom URI's if the meaning is the same. ArthurPSmith (talk) 16:10, 17 October 2018 (UTC)
 - Supposedly you read Lea's announcement about not using others'. Obviously, it would have been easier to this over at Wiktionary, especially I came to the conclude that the French one is actually fairly complete. --- Jura 16:18, 17 October 2018 (UTC)
 - You said this added a "primary mapping for key features of [lexemes] to a scheme that isn't [CC0]". What scheme is not CC0 in the new mapping? As I just linked, ontolex is definitely CC0. ArthurPSmith (talk) 17:21, 17 October 2018 (UTC)
 - @Smalyshev (WMF): forgot to ping you. Can we go ahead and change this. If we do it now, it's still fairly easy to update things. --- Jura 10:49, 18 October 2018 (UTC)
- This is great! I created a page using the Wikidata list template to automatically do a few stats: Wikidata:Lexicographical data/Statistics/AutoGenerated. ArthurPSmith (talk) 14:15, 16 October 2018 (UTC)
- @Lea Lacroix (WMDE): Hmm to Jura's point just above, when I query using wikibase:Lexeme or wikibase:Form I get nothing, but using the ontolex types I find everything. It looks like the export doesn't quite match what is stated in mw:Extension:WikibaseLexeme/RDF mapping? ArthurPSmith (talk) 14:45, 16 October 2018 (UTC)
 - @ArthurPSmith: this is intentional, for performance reasons we only keep one class. The dump and RDF export have both. Smalyshev (WMF) (talk) 17:55, 16 October 2018 (UTC)
 - Ah, ok I guess that's fine. I'm running into an issue with query timeouts in lexemes, not sure why it should be happening since we don't really have many yet do you want to hear about it?... ArthurPSmith (talk) 17:58, 16 October 2018 (UTC)
 - And whatever the performance problem was seems to have resolved or maybe I just changed the query enough to get it to work now, but it's quite fast. I've added a number of specific examples to the Wikidata:Lexicographical data/Ideas of queries page.

Senses are now part of Lexicographical Data

Hello all,

We now have Senses :)

Senses will allow you to describe, for each Lexeme, the different meanings of the word. By using multilingual glosses, very short phrase giving an idea of the meaning. In addition, each of these Senses can have statements to indicate synonyms, antonyms, refers-to-concept and more. By connecting Senses to other Senses and to Items, you will be able to describe precisely the meaning of words with structured and linked data. But the most important thing is that Senses will be able to do is collect translations of words between languages.

Feel free to try editing Senses, and let us know if you have questions or find bugs.

Note: there are still issues with sorting the IDs of Senses, Forms and sorting the glosses, that will be solved later this week. Thanks for your understanding.

Cheers, Lea Lacroix (WMDE) (talk) 10:15, 18 October 2018 (UTC)

Note: Senses will appear on Lexemes during the next few minutes, in the order of L-IDs. We just passed administratif (L19000) :) Lea Lacroix (WMDE) (talk) 10:30, 18 October 2018 (UTC)

I've tried senses on tour (L2330), tour (L2331), tour (L2332). I just added 1 or 2 obvious and trivial senses to these lexemes right now, I'll add more later (these 3 words are *very* polysemic in French, the first might have 10 or 20 senses in the end).

More globally, where are we for structure and properties specific to senses? (I must admit I don't care much about senses so I didn't follow it closely) Cheers, VIGNERON (talk) 10:40, 18 October 2018 (UTC)

There seem to be the most interesting one to me : item for this sense (P5137) 🛶 , although the naming is terrible :) Ideally senses could consist of only one statement with this property, see the pictures in « tour(fr) » for example, they are redundant with the one on the « tour » item. author TomT0m / talk page 10:52, 18 October 2018 (UTC)

Congrats to the team :) one big and important step for the project. author TomT0m / talk page 10:52, 18 October 2018 (UTC)

Retrieved from "https://www.wikidata.org/w/index.php?title=Wikidata_talk:Lexicographical_data&oldid=766865561"

This page was last edited on 18 October 2018, at 10:52.

All structured data from the main, property and lexeme namespaces is available under the Creative Commons CC0 License; text in the other namespaces is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy.