Ontolex-Lemon in Wikidata and other Wikibase instances

David Lindemann
UPV/EHU University of the Basque Country
Unibertsitateko Ibilbidea, 5
01006 Vitoria-Gasteiz
david.lindemann@ehu.eus

Abstract

This paper provides insight into how the core elements of the Ontolex-Lemon model are integrated in the *Wikibase Ontology*, the data model fundamental to any instance of the Wikibase software. This includes *Wikidata lexemes*, which today is probably the largest Ontolex-Lemon use case, a dataset collaboratively built by the community of Wikidata users. We describe how lexical entries are modeled on a Wikibase, including the linguistic description of lexemes, the linking of lexical entries, lexical senses and lexical forms across resources, and links across the domain of lexemes and the ontological part of a Wikibase knowledge graph. Our aim is to present Wikibase as a solution for storing and collaboratively editing lexical data following Semantic Web standards, and to identify relevant research questions to be addressed in future work.

1 Introduction

Wikibase,[1] a set of extensions to MediaWiki[2], is a software solution for storing, collaboratively editing and exhibiting structured data on the web, in the shape of a knowledge graph. The software is used, first and foremost, by Wikidata (Vrandečić and Krötzsch, 2014; Erxleben et al., 2014).[3] Many other instances of the software have emerged since the software packages (and hosting solutions) are freely available.[4] The types of entities described in a Wikibase include *items*, which represent all kinds of real-world objects and ontological concepts, and *lexemes*, describing words. As it will be explained in section 2.1, in a Wikibase, lexemes are described following the core of the Ontolex-Lemon model (McCrae et al., 2017), and so they are on Wikidata, which is today probably the largest open and collaboratively editable Ontolex-Lemon use case. By May 2025, Wikidata described 1.42 Million lexical entries in 1,379 different languages, which compared to earlier figures (Nielsen, 2020) means an exponential growth. German, Russian, Danish, Estonian, English and Malayalam are, in this order, the languages with the most described lexemes. The potential uses of the linguistic descriptions contained in this growing resource, which we present in more detail in section 2.2, are manifold.

Section 2.3 is devoted to Wikibase as a linking hub for descriptions of lexical *entries*, *senses* and *forms* across different resources; cross-resource links are encoded as *external ID* properties.[5] External identifiers do not only constitute hyperlinks for a user to jump between the resources presented in different web portals, but also enable SPARQL federation,[6] that is, a Wikibase's content may be integrated also with the lexical and ontological content of any other Wikibase, including Wikidata, or an RDF database of other kind.[7] In this regard, it is important to point out how RDF is used on a Wikibase, which will be explained in section 2.4. Related to this, an important and distinguishing feature of Wikibase is its multi-layer integration of lexical and ontological entities inside the same database, which will be discussed in detail in section 2.5.

While the lexemes collection on Wikidata is open for continued enrichment, some use cases may require a separate Wikibase instance, in a first phase of a contribution project, or even for a whole project lifetime: A language may be already described on Wikidata, so that any addition would involve lexeme, sense and form disambiguation

[1] See https://wikiba.se.

[2] See https://mediawiki.org.

[3] See https://www.wikidata.org.

[4] See https://wikiba.se/showcase/ and https://wikibase.world, a catalogue of Wikibase instances.

[5] See https://www.wikidata.org/wiki/Wikidata:External_identifiers.

[6] See https://www.w3.org/TR/sparql11-federated-query/.

[7] See https://www.mediawiki.org/wiki/Wikibase/Federation.

and deduplication tasks, which when working on an own instance could be left to the final phase of a project. Also, a dataset to be curated may be regarded too noisy (e. g. in a legacy dictionary digitisation project), or too esoteric (e. g. when dealing with dialectal or historical language data) to be included in Wikidata without previously ensuring relevance and quality. In addition, to work on an own Wikibase instance means freedom in data modeling and community management. Licensing might also be an issue, since Wikidata content is obligatorily licensed according to CC0. Exploring these potentials, we briefly discuss the Wikibase ecosystem for lexeme descriptions in section 2.6.

In the closing section 3, we provide an outlook for open research questions to be worked on, focusing on the relation between the fine-grained modeling proposals made by the Ontolex Community, on the one hand, and the conventions emerging in the community working on Wikidata lexemes, on the other.

2 Ontolex-Lemon on Wikibase

2.1 Lemon core classes

The core of Ontolex-Lemon,[8] that is, the classes `ontolex:LexicalEntry`, `ontolex:LexicalSense`, and `ontolex:Form`, is reused in the Wikibase Ontology,[9] the backbone model of any Wikibase instance.

Wikibase treats the lexical entry, with its own numeral identifier preceded by letter L, as primary entity describing a *lexeme*, with forms and senses as sub-entities, and presents the instances of those three Ontolex-Lemon core classes together on one editable lexeme page.[10] This structure is pre-set in the data model fundamental to any Wikibase instance and cannot be modified by the user, and the same is true for a small number of properties to be attached to the three core classes, listed in table 1; the three obligatory properties describing a lexical *entry* must have a value, a *sense* must have a *gloss* (a short sense-disambiguating text, represented in RDF using `skos:definition`), and a *form* must have a representation (a value for `ontolex:representation`).[11]

`ontolex:LexicalEntry`	
	`wikibase:lemma`
	`wikibase:lexicalCategory`
	`dct:language`
	`ontolex:sense`
	`ontolex:lexicalForm`
`ontolex:LexicalSense`	
	`skos:definition`
`ontolex:Form`	
	`wikibase:grammaticalFeature`
	`ontolex:representation`

Table 1: Obligatory basic classes and properties in their domain for describing lexemes in Wikibase

In detail, this obligatory structure entails the following restrictions:

- A lexical *entry* must point to exactly one item in the same Wikibase as value for *lexical category*.
- A lexical *entry* must point to exactly one item in the same Wikibase as value for its *language*.
- A lexical *entry* must have at least one *lemma*. Several lemmata can co-exist for the same *entry* for covering different spelling variants (e. g. British and American English).[12] In the RDF representation, lexeme lemmata appear attached to the entry using a property named `wikibase:lemma`. Lemmata are indexed for the MediaWiki text search index,[13] so that a user can search for lexemes, manually in the interface, or via API.
- A lexical *sense* must have at least one *sense gloss*, in any language, i. e. not necessarily or not only in the language associated to the *entry*. Purpose of the gloss is to provide the information necessary to discriminate word senses.
- A *form* must have at least one *written representation*. Alike different values for *lemma*

[8] For the original Lexical Model for Ontologies, see `https://lemon-model.net/`.

[9] See `https://wikiba.se/ontology/`.

[10] See `https://www.wikidata.org/wiki/Lexeme:L1` for the lexeme page describing lexeme `wd:L1`.

[11] Although `ontolex:writtenRep` would be the best match here, the Wikibase Ontology uses `ontolex:representation`; in Ontolex-Lemon, the former is defined as subproperty of the latter.

[12] See an example at `https://www.wikidata.org/wiki/Lexeme:L1347`, where the English *color/colour* is described, an example for an entry with lemmata in distinct scripts at `https://kurdi.wikibase.cloud/wiki/Lexeme:L3447`.

[13] See `https://www.mediawiki.org/wiki/Elasticsearch`.

on entry level, form representations in several spelling variants can be attached to the same *form* entity. A written representation in most cases will be a string as found in text of that language, but written representations also include code transcriptions, such as those describing sign language forms.[14] To describe a form without providing any type of written representation is not foreseen.

- A *form* can have zero or more items in the same Wikibase as values for *grammatical feature*.

Lemmata, *sense* glosses (skos:definition), and also *form* representations (ontolex:representation) are associated to a language code. The available codes are the same that are available throughout the mediaWiki instance, e. g. for labels, descriptions, and *monolingualtext* strings.[15]

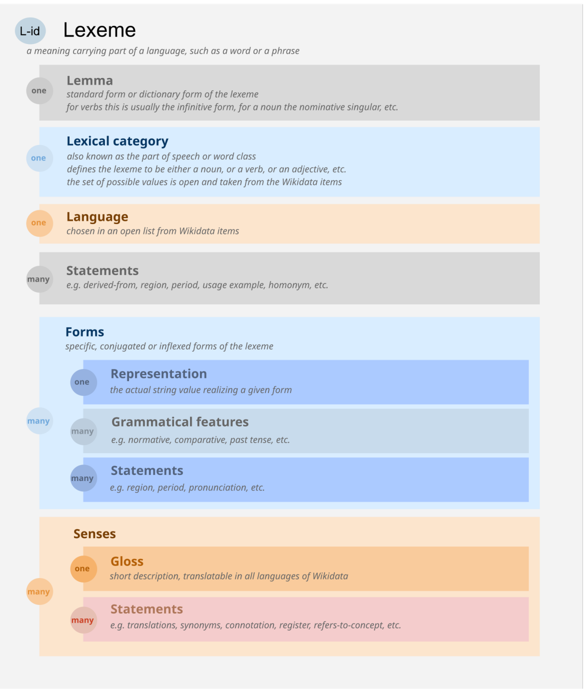
L-id
Lexeme
a meaning carrying part of a language, such as a word or a phrase
one
Lemma
standard form or dictionary form of the lexeme
for verbs this is usually the infinitive form, for a noun the nominative singular, etc.
one
Lexical category
also known as the part of speech or word class
defines the lexeme to be either a noun, or a verb, or an adjective, etc.
the set of possible values is open and taken from the Wikidata items
one
Language
chosen in an open list from Wikidata items
many
Statements
e.g. derived-from, region, period, usage example, homonym, etc.
many
Forms
specific, conjugated or inflexed forms of the lexeme
one
Representation
the actual string value realizing a given form
many
Grammatical features
e.g. normative, comparative, past tense, etc.
many
Statements
e.g. region, period, pronunciation, etc.
many
Senses
one
Gloss
short description, translatable in all languages of Wikidata
many
Statements
e.g. translations, synonyms, connotation, register, refers-to-concept, etc.

Figure 1: The Wikibase lexeme, as illustrated in the Wikibase documentation

These restrictions guarantee interoperability on a basic level. Beyond that, any additional relation involving *entry*, *sense* or *form* objects is not predefined and can be modeled according to the use case using self-defined properties in Wikibase *statements*,[16] as illustrated in Fig. 1.[17] The RDF classes and properties mentioned above are part of the Wikibase Ontology, that is, they are used in the RDF representation of an entity (an *item*, a *lexeme*, a *property*), and accordingly, they appear in RDF entity data dumps.[18] In opposition to that, all other relations added to *lexeme*, *sense* or *form* as Wikibase *statements* (see section 2.4) always involve a property defined in the namespace of the Wikibase instance itself, identified by a number preceded by the letter P, and that has a range restricted to one Wikibase datatype.[19]

2.2 Wikidata lexemes

Wikidata lexemes is an open and editable collection, where everybody is invited to collaborate. A documentation of the data model based on the three Ontolex-Lemon core classes is given on the Wikidata documentation pages.[20] Concerning advanced modeling questions, contributors get support from each other.[21] A core group of more experienced and active users provides advice to newcomers and occasional contributors, also through dedicated outreach events.[22] A manually curated list of lexical entries provides examples in several languages of good and complete modeling practice.[23]

Instead of following prescribed models concerning morphology, etymology, multilingual equivalents, etc., on Wikidata, a bottom-up grown set of properties is used for describing entries, senses, and forms; see table 2 for the ten most frequent properties for each of the classes, pointing to other

[14] For an example, see wd:L991786-F1.

[15] See https://www.wikidata.org/wiki/Help:Monolingual_text_languages, and for a list of all codes implemented in a Wikibase instance, e. g. https://www.wikidata.org/w/api.php?action=query&meta=wbcontentlanguages.

[16] See https://www.wikidata.org/wiki/Help:Statements.

[17] The picture is used at https://www.mediawiki.org/wiki/Extension:WikibaseLexeme/Data_Model.

[18] In Turtle serialization, see, for example, https://www.wikidata.org/wiki/Special:EntityData/L1347.ttl.

[19] See https://www.wikidata.org/wiki/Help:Data_type.

[20] See https://www.mediawiki.org/wiki/Extension:WikibaseLexeme/Data_Model.

[21] Discussions take place on-wiki (see https://www.wikidata.org/wiki/Wikidata_talk:Lexicographical_data) and in a dedicated channel on the *Telegram* platform.

[22] The pages dedicated to the 2021 and the 2024 *Lexicodays* provide video recordings, presentation slides and links to other pages containing lexicographical guidelines and descriptions of tools related to Wikidata lexemes, see https://www.wikidata.org/wiki/Wikidata:Events/Lexicodays_2024 and https://www.wikidata.org/wiki/Wikidata:Events/30_lexic-o-days_2021

[23] See https://www.wikidata.org/wiki/Wikidata:Showcase_lexemes.

entities on Wikidata, or to a data value (in table 2, excluding *external id* properties). Without going into much detail, we point out some of them; in some cases, the modeling is unambiguous (and straightforwardly alignable to Ontolex), and sometimes alternative modeling approaches coexist:

- As it will be explained in section 2.5, **translation equivalence** is expressed using `wd:P5972`, a property set (in both directions) directly between senses; the property used in Ontolex-Lemon for this purpose is `vartrans:translation`. In May 2025, Wikidata contained 119,847 translation links between senses of different languages.

- For representing **etymology**, a lexeme is linked directly to the lexical entry describing the etymon using `wd:P5191`.[24] In May 2025, Wikidata contained 40,540 etymology links from lexeme to lexeme.

- For representing **decomposition**, `wd:P5238`, the *combines lexeme* property links a compound or multiword entry to its constituents, in the same way `decomp:subterm` is used according to Ontolex.[25] This property is used in May 2025 207,799 times.

- **Pronunciation** is described in multiple ways, but always associated to *forms*; the value of a general *pronunciation* property `wd:P7243` is in general identical to the form representation, but may include stress indicators and other diacritics, e. g. to indicate vowel length. It is recommended that pronunciation audio files and/or IPA transcriptions are attached as *qualifiers* to the pronunciation claim, but audio files can also be directly attached to *form*.[26] Today, 74% of 252,652 pronunciation audio files are linked directly to a *form*. Only audio files hosted on Wikimedia Commons can be used.[27]

- As seen in table 2, properties devoted to certain **transcription** and **transliteration** systems exist; each of them is, like all Wikidata properties, described through its entity data, and on an own discussion page.[28] In parallel, a general property *transliteration or transcription* (`wd:P2440`) can be used, and its value qualified using the property *determination method or standard* (`wd:P459`). The former, defined as subproperties of `wd:P2440`, are heavily used on Wikidata in general,[29] but in only about 1.2% of its use cases (i. e., not more than 3,808), that property is attached to *forms*,[30] on which the general property is used around ten times more often: The 70,165 uses of `wd:P2440`, with a `wd:P459` qualifier pointing to the transliteration system, are almost half-divided between *items* and *forms*.[31]

- **Usage examples** (`wd:P5831`) are recommended to be attached to *entry*, and not to *sense*. However, about 7% of the 31,271 usage examples in Wikidata lexemes today remain attached to *sense*; the distribution across languages shows a diverse picture, but a clear preference for *entry* as subject of the property.[32] If attached to *entry*, one or more senses can be declared *subject sense* (`wd:P6072`); this is done using that property as qualifier in the example statement. The advantage of this modeling, apart from being able to declare a usage example to be pertinent to more than one sense, lies in the ability to have examples attached to the entry also if in the moment of upload and without or before whatever sense disambiguation procedure it is not clear which sense should be marked as *subject sense*, e. g. when dealing with examples stemming from corpora, or if (still) no senses are described for the lexeme: As soon as the correct sense can be determined, the `wd:P5831` *claim* is enriched with with a `wd:P6072` *qualifier*, without having to delete and re-write the whole *statement* (with its references), and attach it to *sense*. Another strong reason for attaching

[24] See, for example, `https://www.wikidata.org/wiki/Lexeme:L630740#P5191`.

[25] See `wd:L625224` for the German phrase "frohes neues Jahr".

[26] For an example, see `wd:L3338-F2`. *Qualifiers* as part of a *statement* are explained in section2.4

[27] The property is of datatype *Commons Media File*, see `https://www.wikidata.org/wiki/Help:Data_type#commonsMedia`.

[28] For example, about the property `wd:P5825` *ISO 15919 transliteration*, a documentation is available at `https://www.wikidata.org/wiki/Property_talk:P5825`, and about `wd:P4187` *Tibetan to Latin transliteration*, `https://www.wikidata.org/wiki/Property_talk:P4187`.

[29] See `https://w.wiki/EBWF`.

[30] See `https://w.wiki/EBYi`.

[31] See `https://w.wiki/EBZ2` for usage counts according to the different transliteration systems.

[32] See `https://w.wiki/ECBm` for the use of this property across languages.

usage examples to entry is to enable their annotation with a subject *form*, i. e. the word form appearing in the example.[33] In addition, having examples at both levels complicates their retrieval in queries.

The Wikidata lexemes collection is quantitatively described on dedicated pages,[34] and it can be explored using the *Ordia* tool (Nielsen, 2019),[35] which generates dictionary-like exhibitions of lexical data; it also features a tool to look up Wikidata *forms* matching to tokens in text. *Ordia* also provides statistics on the Wikidata lexeme collection, such as, for example, lists of the most frequently used properties in the domains of *entry* (a. k. a. *lexeme*), *sense* and *form*.[36] The *Synia* tool also shows statistics on lexemes, e. g. counts of values for `wd:P6191` *language style* in different languages.[37]

Table 3 lists overall counts for the ten languages with best absolute coverage at the three levels.[38] Asking for relative coverages, without data about the total amount of corpus lemmata, corpus types, and dictionary senses in a language, we might ask for a relation between the coverage on Wikidata and the number of speakers of a language. A query like that (taking into account languages with more than 100,000 speakers) results in a different ranking, with Estonian, Breton, Basque and Danish leading lexemes,[39] and Basque, Breton, Dagbani and Norwegian Bokmål as top four languages for senses.[40]

LexicalEntry	
`wd:P5185`	*grammatical gender*
`wd:P5911`	*paradigm class*
`wd:P5238`	*combines lexeme*
`wd:P31`	*instance of*
`wd:P5187`	*word stem*
`wd:P1552`	*has characteristic*
`wd:P5402`	*homograph lexeme*
`wd:P2348`	*time period*
`wd:P5191`	*derived from lexeme*
`wd:P5186`	*conjugation class*
`wd:P5831`	*usage example*
`wd:P7486`	*grammatical aspect*
LexicalSense	
`wd:P5137`	*item for this sense*
`wd:P5972`	*translation*
`wd:P5973`	*synonym*
`wd:P1343`	*described by source*
`wd:P18`	*image*
`wd:P9488`	*field of usage*
`wd:P6191`	*language style*
`wd:P8394`	*gloss quote*
`wd:P9970`	*predicate for*
`wd:P6271`	*demonym of*
`wd:P6084`	*location of sense usage*
`wd:P10339`	*semantic gender*
Form	
`wd:P7243`	*pronunciation*
`wd:P898`	*IPA transcription*
`wd:P443`	*pronunciation audio*
`wd:P5279`	*hyphenation*
`wd:P5825`	*ISO 15919 transliteration*
`wd:P8881`	*ITRANS (Indic scripts)*
`wd:P8530`	*alternative form*
`wd:P5276`	*Slavistic Phonet. Alphab. transcr.*
`wd:P10822`	*homophone form*
`wd:P1721`	*Hanyu Pinyin transliteration*
`wd:P11950`	*appears before phonolog. feat.*
`wd:P11951`	*appears after phonolog. feat.*

Table 2: The 12 most frequently used properties describing Wikidata lexemes, senses and forms (May 2025)

Apart from those already mentioned, a range of tools[41] has emerged around Wikidata lexemes, designed to help creating entries and append senses,[42] for linking senses to ontological references,[43] for creating forms collections through templates,[44] searching for usage examples and adding them to

[33] See, for example, the usage examples for `wd:L87`.

[34] See `https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Statistics`

[35] Accessible at `https://ordia.toolforge.org/`.

[36] Accessible at `https://ordia.toolforge.org/property/`. Wikidata SPARQL queries as used in *Ordia* can be modified for custom searches, e. g. for listing the property statistics for a single language, as in the following queries derived from those in *Ordia*: *entry*, `https://w.wiki/E4pg`, *sense*, `https://w.wiki/E4pw`, *Form*, `https://w.wiki/E4p$`; note the filter for *external id* properties.

[37] See `https://synia.toolforge.org/#languagestyle`.

[38] See up-to-date statistics at `https://w.wiki/ECLq` for *lexeme*, at `https://w.wiki/EDkc` for *sense*, and at `https://w.wiki/EDjh` for *form*.

[39] See `https://w.wiki/EDna`.

[40] See `https://w.wiki/EDnQ`.

[41] See also `https://www.wikidata.org/wiki/Wikidata:Tools/Lexicographical_data`.

[42] For a lexeme creation UI, see `https://hangor.toolforge.org`; for python, in addition to the general *wikibaseintegrator* library (`https://github.com/LeMyst/WikibaseIntegrator`), specially for lexemes, *tfsl* (`https://gitlab.wikimedia.org/toolforge-repos/twofivesixlex`.)

[43] See `https://lexica-tool.toolforge.org/`.

[44] See `https://www.wikidata.org/wiki/Wikidata:Wikidata_Lexeme_Forms`.

LexicalEntry	language	**LexicalSense**	language	**Form**	language
1,422,331	all Wikidata	**585,893**	all Wikidata	**14,396,207**	all Wikidata
239,465	German	48,308	Bokmål	2,802,106	Estonian
102,150	Russian	41,973	English	1,257,083	Basque
96,753	Danish	30,756	Basque	1,246,745	Russian
83,218	Estonian	29,790	Nynorsk	1,199,194	Latin
68,758	English	24,127	Czech	871,813	Czech
67,367	Malayalam	24,055	Swedish	753,118	Malayalam
64,445	Italian	22,000	Italian	692,671	Spanish
63,128	Spanish	21,036	Japanese	641,454	Danish
56,296	Latin	20,484	Persian	571,091	German
48,214	Swedish	18,851	Egyptian	522,107	Italian

Table 3: Absolute coverage of some languages in Wikidata lexemes (May 2025)

entries,[45] for massively recording pronunciation audios,[46] or for graph visualisations of lexical relations.[47]

One goal of the Wikidata lexemes collection is to enable natural language generation for drafting Wikipedia article text from abstract knowledge representations (Vrandečić, 2021; Morshed, 2024),[48] and another possible application is corpus annotation (Lindemann and Alonso, 2024); all these depend on the degree the lexemes, senses and forms in the collection cover the languages to process.

2.3 Wikibase lexemes as linking hub: The case of Wikidata

In addition to these and other properties for the linguistic description of the lexeme, Wikibase lexical entries contain pointers to external resources encoded as *external id* properties. These lead a human user to an entry or sense description in a dictionary web portal. In some cases, federated database queries can access content in several graph databases at a time using such external ID. For example, a query can involve Wikidata and the LiLa Latin Knowledge Base (Passarotti and Mambrini, 2022),[49] exploiting the LiLa URI attached to Wikidata Latin lexemes, and calling the LiLa SPARQL endpoint from within the Wikidata Query Service, or vice versa.

Wikidata lexemes, on *entry* level, by May 2025 count 2.3 Million *external id* statements. Table 4 shows usage counts for the most frequent 20 properties.[50] As for *sense*, all lexical resources aligned to Wikidata on sense level today still lack significant coverage,[51] although we point out the fact that e. g. English Wordnet synset identifiers are aligned to Wikidata *items* (McCrae and Cillessen, 2021), which, as explained, are referenced by a significant number or senses. An alignment of *forms* to external resources such as corpus-based form repositories is at large still not present, although it would be interesting, for instance in rich-morphology languages, where the morphologically possible forms outnumber the forms that are actually attested in corpora, so that forms attestation is very valuable information, with a similar value lemma attestation has in languages with a comparably reduced number of different inflected forms, like English.

2.4 Reification in Wikibase

Wikibase *statements* include by default a mechanism for further describing the main *claim* of a statement using *qualifiers*, *ranks*, and *references*. A graphical model of a Wikibase statement is given in figure 2, where "entity" represents an *item*, *lexeme*, *sense*, *form*, or *property* node, each of which has its own URI in the main entity namespace of the Wikibase instance,[52] and where the blue-colored "value" nodes, depending on the property datatype, represent entities of the same Wikibase, or data val-

[45] The *Luthor* tool uses Wikisource content as corpus, see https://luthor.toolforge.org/.

[46] See https://lingualibre.org/.

[47] For an example, an etymological network, see https://lucaswerkmeister.github.io/wikidata-lexeme-graph-builder/?subjects=L184995&predicates=P5191.

[48] See https://meta.wikimedia.org/wiki/Abstract_Wikipedia.

[49] See https://lila-erc.eu; the LiLa ID property is wd:P11033.

[50] For up-to-date counts, see https://w.wiki/E9j3.

[51] An example for one of the most used external identifiers is *DWDS sense ID*, see https://www.wikidata.org/wiki/Property_talk:P12550.

[52] On Wikidata, e. g. wd:Q1 for an item, wd:L1 for a lexeme, wd:L1-S1 for a sense, wd:L1-F1 for a form, and wd:P31 for a property.

Count	Property	Property label (en)	Lang. (ISO-639-1)
231,859	wd:P9940	*DWDS-Lemma-Identifikator*	de
153,322	wd:P8376	*Duden-Lexem-Identifikator*	de
139,067	wd:P11519	*elexiko ID*	de
122,350	wd:P11138	*Sõnaveeb entry ID*	(45 lang.)
84,951	wd:P9947	*WDG-Lemma-Identifikator*	de
66,512	wd:P13258	*Presisov večjezični slovar ID*	fr de en sq sh
65,599	wd:P9529	*Den Danske Ordbog article ID*	da
61,333	wd:P5912	*Oqaasileriffik online dictionary ID*	da en kl nb
54,755	wd:P12630	*Aragonario ID (6th version)*	es an
51,566	wd:P11033	*LiLa Linking Latin URI*	la
49,987	wd:P9385	*DWB Lemma ID*	de
45,751	wd:P5275	*OED Online ID*	en
39,052	wd:P9962	*Ordbog over det danske sprog ID*	da
37,925	wd:P12420	*Il Nuovo De Mauro ID*	it
37,137	wd:P10042	*Bokmålsordboka-ID*	nb nn
36,535	wd:P11838	*Svenska Akademiens ordlista ID*	sv
35,124	wd:P9387	*GWB Lemma ID*	de
31,014	wd:P12690	*New Oxford American Dictionary ID*	en
29,803	wd:P11319	*Little Academic Dictionary ID*	ru
29,316	wd:P12828	*DAKA Danish-Greenlandic Dictionary ID*	da

Table 4: Most frequently used *external id* properties on Wikidata lexemes, and lexeme languages

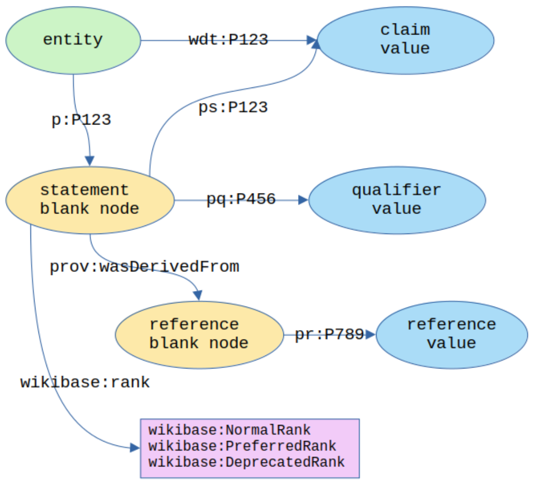

Figure 2: Graphical model of a Wikibase Statement

ues, including strings, external identifiers, date objects, globe coordinates, et cetera. Each statement node may be linked to several *qualifier* values, and to several *reference* nodes (in the RDF representation attached to the statement blank node using prov:wasDerivedFrom), which make up blocks of references. *Ranks* are used for annotating multiple statements for the same property with one of the three values *normal*, *preferred*, or *deprecated*.[53]

In order to keep all data pertaining to the same lexical entry "together", so that it would all get stored in the same entity data JSON blob,[54] and displayed on the same lexeme entity page, it is convenient to keep the modeling of the linguistic description as shallow as a list of Wikibase *statements*. Unlike in the Ontolex modules, where the linguistic description often involves several reification layers, the model followed in Wikidata, and, for the same reason, in any other Wikibase, will try, wherever possible, to stay with certain reification approaches, which can be combined, but will be limited to the following:

- Using statement *qualifiers*, i. e. semantic triples describing the main *claim* of the statement.

- Using *references* for provenance annotations.

- Using subproperties such as wd:P1721

[53] See https://www.wikidata.org/wiki/Help:Ranking.

[54] At https://www.wikidata.org/wiki/Special:EntityData/L1.json, the entity data JSON for wd:L1 as stored in the database, and from which the entity page display is generated, can be obtained.

Hanyu Pinyin transliteration, subproperty of `wd:P2440` *transliteration or transcription*.

If we compare, for example, how a translation relation between two senses can be further described, according to Ontolex this can be modeled introducing a blank node of class `vartrans:Translation` into the dataset (option A), linking that to both *sense* nodes using typed properties (*source* and *target*), and attaching to that node supplemental information about the translation relation (Bosque-Gil et al., 2015). As a shallower alternative (option B), which does not involve any additional blank node in the structure, and consequently allows no additional description of the translation relation, Ontolex uses the `vartrans:translation` property.[55] The second option is more suitable for a Wikibase, because the additional node in option A, since it does not fit into the Wikibase statement structure, would have to be created as named individual entity, i. e. a Wikibase *item* with its own Q-identifier, its labels to be indexed in the Wikibase ElasticSearch, its class declaration, and descriptions. However, in a Wikibase, option B caters for a description as rich as option A in Ontolex: In the discussed example, the translation relation type, its direction, or any other translation restriction feature can be expressed using a *qualifier* on the translation *claim*. This, in turn, is not possible when using option B in Ontolex, since the semantic triple describing the translation relation there cannot be further described or qualified.

2.5 Lexicon-Ontology interface and multilinguality

According to Ontolex-Lemon, a property named `ontolex:reference` links a lexical sense to an ontological concept (Bosque-Gil et al., 2015). On Wikidata, `wd:P5137` *item for this sense* has the equivalent function. For example, Wikidata's lexeme `wd:L3549`, describing the English noun *foot*, has three senses, each of them pointing to a different conceptual *item* node in the graph, using that property. `wd:L3549-S1`, the first listed sense, is linked to an item describing a unit of length, while the second sense points to an item describing a furniture part, and the third sense, `wd:L3549-S3`, glossed as "anatomical structure found in vertebrates", links to the anatomical entity. Each of the three linked Wikidata *items* is itself further described, for example, with links to Wikipedia articles in multiple languages (the property used here is `schema:about`), which have a title in that language, and which provide encyclopaedical descriptions of the concept. That means in general terms that Wikibase provides a framework where lexical and ontological (conceptual) descriptions converge, and where text pages (for Wikidata, Wikipedia articles) *about* concepts also have their habitat. Since, in addition, Wikibase items are annotated with multilingual labels (`rdfs:label` and `skos:altLabel`) and descriptions (`schema:description`), the `wd:P5137` reference of a lexeme sense into the ontological part of the Wikidata graph already provides three facets of multilinguality: The labels, textual concept descriptions, and entire text pages attached to an ontological item referenced by lexeme senses may cover multiple languages.

Since several languages' word senses can be linked to the same *item*, `wd:P5137` provides translation (and, inside the same language, synonymy) information. In Ontolex, this way of modeling translation relations is referred to as *translation as shared reference*.[56] As of May 2025, Wikidata contains 227,908 *item for this sense* claims that link senses to items.[57] Calculating the number of translation links through shared references for every item, that sums 5.23 million translation (and intralingual synonymy) links between lexeme senses, counting the connecting links twice, i. e. as translation link in both directions.[58] This by far outnumbers `wd:P5972` *sense translation* relations (see section 2.2), which constitute an alternative without leaving the domain of lexemes, as needed for senses without an ontological reference in Wikidata (most prominently, senses of lexemes with a lexical category other than *noun*). As described above, multilingual sense *glosses* (very short sense descriptions attached using the built-in `skos:definition`) provide another facet of multilingual sense description.

Fig. 3 shows an example for a lexeme's relations in the graph, including *monolingualtext* values in two languages: the lexical entry describing a German noun *Pferd* ("horse"), linked to entities of dif-

[55] See examples with figures for both at `https://www.w3.org/2016/05/ontolex/#translation-as-a-relation-between-lexical-senses`.

[56] See `https://www.w3.org/2016/05/ontolex/#translation-as-shared-reference`.

[57] See the distribution of *item for this sense* across languages at `https://w.wiki/ECCA`.

[58] See `https://w.wiki/EcDh` for a list of all items linked to from senses, and the corresponding number of translation links (includes intralingual translation, i. e. synonymy).

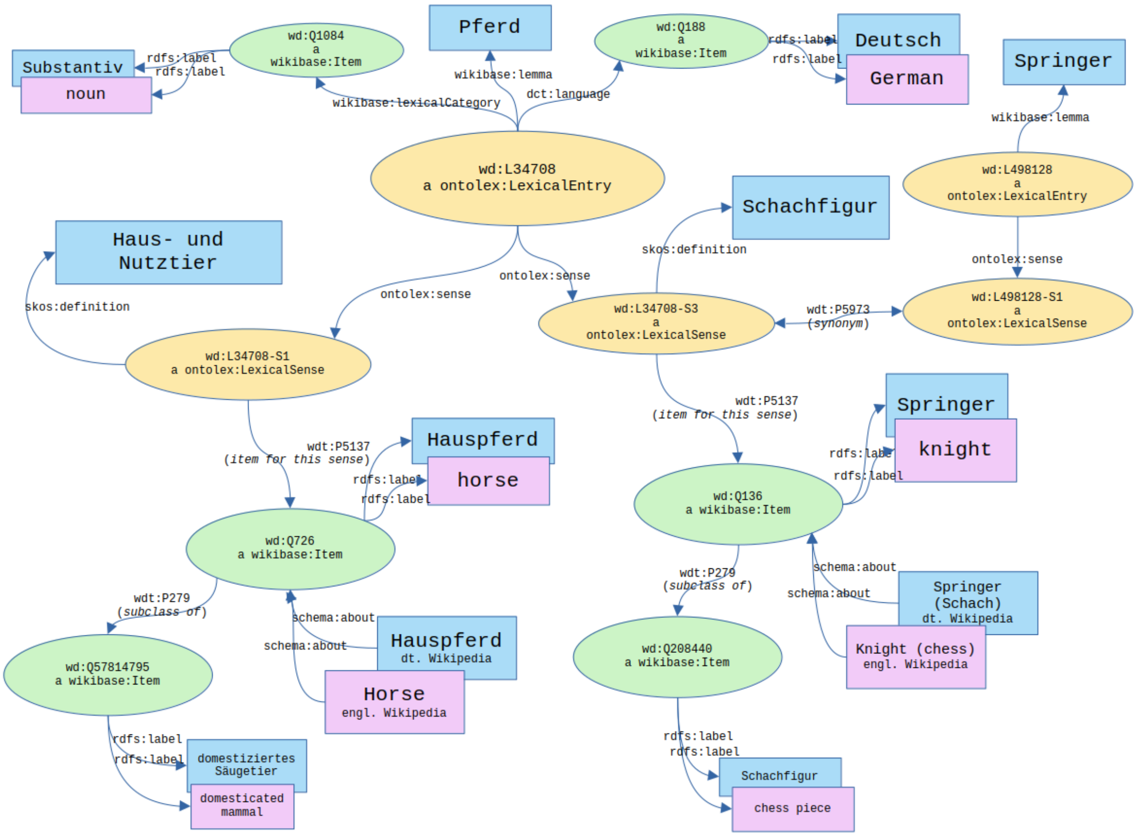

Figure 3: Wikidata entry wd:L34708 describing the German lemma *Pferd* and some relations

ferent types: ontological concepts (*items*), other lexical entries (*lexemes*), and lexical senses. Entities of type *item* are used to represent the *language*, which allows querying for lexemes according to language features such as the language family or countries where languages are spoken or native to, to represent the *lexical category*,[59] and, as explained above, for ontological sense references. *Label*, *definition* and *lemma* values are always strings associated to a language code (figure 3 only shows English and German, while the cited entities on Wikidata cover more languages here).

2.6 Wikibase as an infrastructure for lexical datasets

In terms of the *FAIR Guiding Principles for scientific data management and stewardship*[60], a lexical dataset on a Wikibase can safely be called to be state of the art, since permanent URI on the level of the three Ontolex-Lemon core classes assure *findability*, answering human user calls with the display of an editable entity page, and programmatical calls with machine-readable data in various formats.[61] *Accessibility* is furthermore given through the graphical query service,[62] and through a SPARQL endpoint. *Interoperability* is sustained by re-using W3C-recommended RDF vocabularies for a set of basic classes and properties defined in the Wikibase Ontology, such as Ontolex-Lemon for lexical data. Finally, *reusability* is ensured by open licenses.[63]

Wikimedia Deutschland, the WMF chapter responsible for Wikidata, is providing the Wikibase software for self-hosting,[64] and also provides a

[59] On Wikidata, a range of 320 different *items* is used here, some defined as subclasses of others, e. g. wd:Q1166153 *intransitive verb subclass of* wd:Q24905 *verb*; see https://w.wiki/ECKP for use counts.

[60] See https://www.go-fair.org/fair-principles/

[61] Entity data dumps are available in JSON and TTL format.

[62] For Wikidata, accessible at https://query.wikidata.org/.

[63] Wikidata declares its terms of use at the bottom of every displayed page, and in detail at https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use, and any other Wikibase may contain declarations in a similar form.

[64] See https://www.mediawiki.org/wiki/Wikibase/Docker.

hosting cloud.[65] In community discussions, the vision connected to that embraces an ecosystem of independent but federated Wikibases, with Wikidata as central linking hub.[66] Regarding several domains of knowledge, this is already becoming reality; also related to lexical data, some projects have been able to showcase the potential this infrastructure provides for interlinked, FAIR datasets, in experiments with lexical datasets derived from different kinds of sources, such as dictionaries in CSV format (Huaman et al., 2023), Ontolex-Lemon TTL (Lindemann et al., 2023), and, recently, DMLEX (Krek et al., in press).

3 Outlook to further research

In this paper, we have been revisiting the model for lexical entries on Wikibase, and in Wikidata lexemes, the largest collection of lexical data on a Wikibase today, and which is also, probably, the largest Ontolex use case. It can be stated that, by reusing Ontolex-Lemon, the Wikibase software enables the community to perform steps towards the vision of a *Linked Lexical Data Cloud* (Declerck, 2018).

Since the first publications of the Lemon model (McCrae et al., 2012), which had been available by the time of defining the Wikibase Ontology, the Ontolex community has been publishing modeling proposals as modules, regarding, among other aspects, the description of morphology, etymology, and corpus attestations.[67] In the same timespan and in parallel, there has been emerging a tradition of modeling lexemes on Wikidata. One research question seems obvious in this context: Where and how do both models differ, where do they come together? Where can they benefit from each other? What advantages and disadvantages have the two strategies, compared to each other: A top-down model with strict recommendations, aiming at interoperable lexical datasets also at a more fine-grained level (Ontolex), and a model that limits obligatory interoperability to the core, leaving decisions regarding fine-grained descriptions up to the user (Wikibase), aiming at higher levels of interoperability through on-the-fly grassroot community discussions (Wikidata)? This can shed lights on questions about whether the Lemon core has proven its functionality, if the obligatory core should be extended, or if even some of the minimum requirements might turn out problematic for certain use cases. Comparing both universes may lead to useful insight: Can the Wikidata lexemes collection provide data-driven evidence for modeling decisions that can be regarded universal, beyond the Ontolex-lemon core? And, in the other direction: Can collaborators or automatic tools informed in Ontolex-Lemon modeling proposals help grassroot communities to improve consistency, quality assessment and interoperability? A continued dialogue between the Ontolex and the Wikidata lexemes communities, and those around other instances of the Wikibase software, will help to address these questions in detail.

Acknowledgements

The research leading to the presented results has been supported by DLTB research group (Basque Government, IT1534-2).

References

Julia Bosque-Gil, Jorge Gracia, Guadalupe Aguado-de Cea, and Elena Montiel-Ponsoda. 2015. Applying the OntoLex Model to a Multilingual Terminological Resource. In *The Semantic Web: ESWC 2015 Satellite Events*, pages 283–294.

Thierry Declerck. 2018. Towards a Linked Lexical Data Cloud based on OntoLex-Lemon. In *Proceedings of the LREC 2018 Workshop "6th Workshop on Linked Data in Linguistics LDL-2018"*, Miyazaki, Japan.

Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. Introducing Wikidata to the Linked Data Web. In Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble, editors, *The Semantic Web – ISWC 2014*, number 8796 in Lecture Notes in Computer Science, pages 50–65. Springer International Publishing, Cham.

Elwin Huaman, David Lindemann, Valeria Caruso, and Jorge Luis Huaman. 2023. QICHWABASE: A Quechua Language and Knowledge Base for Quechua Communities.

Simon Krek, Primož Ponikvar, Andraž Repar, Iztok Kosem, and David Lindemann. in press. DMLEX on Wikibase: Legacy dictionaries as collaboratively editable dataset. In *Proceedings of eLex 2025: Electronic Lexicography in the 21st Century - Intelligent Lexicography*, Bled.

[65] See https://wikibase.cloud.

[66] See https://meta.wikimedia.org/wiki/LinkedOpenData/Strategy2021/Joint_Vision.

[67] See an overview and source data at https://github.com/ontolex/ontolex.

David Lindemann, Sina Ahmadi, Anas Fahad Khan, Francesco Mambrini, Federicia Iurescia, and Marco Carlo Passarotti. 2023. When OntoLex Meets Wikibase: Remodeling Use Cases. *CEUR Workshop proceedings*, 2773.

David Lindemann and Mikel Alonso. 2024. Linking Historical Corpus Data and Annotations using Wikibase. In Kristina Štrkalj Despot, Ana Ostroški Anić, and Ivana Brač, editors, *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*, pages 743–748. Institute for the Croatian Language, Zagreb.

John Philip McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. 2012. The lemon cookbook.

John Philip McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: Development and Applications. In *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017*, pages 587–597, Brno. Lexical Computing CZ s.r.o.

John Philip McCrae and David Cillessen. 2021. Towards a Linking Between Wordnet and Wikidata. In *Proceedings of the 11th Global Wordnet Conference*, pages 252–257.

Mahir Morshed. 2024. Using Wikidata lexemes and items to generate text from abstract representations. *Semantic Web*, 16.

Finn Årup Nielsen. 2019. Ordia: A Web Application for Wikidata Lexemes. In Pascal Hitzler, Sabrina Kirrane, Olaf Hartig, Victor De Boer, Maria-Esther Vidal, Maria Maleshkova, Stefan Schlobach, Karl Hammar, Nelia Lasierra, Steffen Stadtmüller, Katja Hose, and Ruben Verborgh, editors, *The Semantic Web: ESWC 2019 Satellite Events*, volume 11762. Springer, Cham.

Finn Årup Nielsen. 2020. Lexemes in Wikidata: 2020 status. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 82–86, Marseille, France. European Language Resources Association.

Marco Carlo Passarotti and Francesco Mambrini. 2022. Linking Latin: Interoperable Lexical Resources in the LiLa Project. In *Building new resources for historical linguistics*, pages 103–124. Pavia University Press, Pavia.

Denny Vrandečić. 2021. Building a multilingual Wikipedia. *Communications of the ACM*, 64(4):38–41.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.