Extension:WikibaseLexeme/Data Model

< Extension:WikibaseLexeme

This is a living document, describing the conceptual data model used by WikibaseLexeme. It is not a specification of any concrete binding, implementation, mapping, or serialization.

Lexeme:
- Lemma
- Language
- Lexical category
- Statements
- Forms:
  - Representation
  - Grammatical Features
  - Statements
- Senses:
  - Gloss
  - Statements

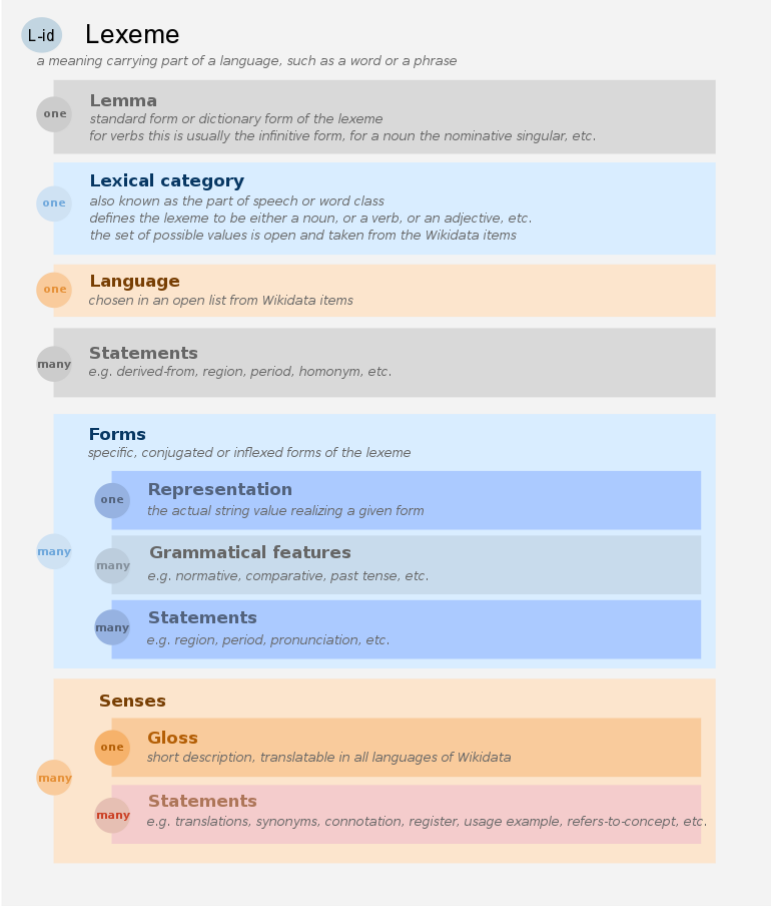
The **data model of WikibaseLexeme** describes the structure of the data that is handled as "Lexemes" in Wikibase, such as words and phrases. While it would be theoretically possible to model these things using Items, a more expressive specialized model helps to reduce complexity, and improve re-use and mappings to other vocabularies. This data model is conceptual ("Which information do we have to support?") and does not specify how this data should be represented technically ("Which data structures should the software use?") or syntactically ("How should the data be expressed in a file?"). Separate documents describe the serialization of the Wikibase data model in JSON and in RDF (Resource Description Framework). The Lexeme data model defines basic concepts and relationships needed to describe lexemes, which act as a fixed ontology. This ontology provides a minimal scaffolding that allows Items and Statements to be used for detailed modeling of a lexeme. The specification of the Lexeme data model is based on the Wikibase data model, so the Wikidata glossary and the Wikibase data model primer may be helpful in understanding this document. The Lexeme data model aims to align with the LEMON model (https://www.w3.org/2016/05/ontolex/) by the Ontolex (https://www.w3.org/community/ontolex/) W3C community group, where useful and practical. However, in the spirit of Wikibase, the Lexeme model is designed to be simple and flexible enough for casual collaborative editing, as opposed to the more formalized approach taken by LEMON.

Lexeme

A Lexeme is a lexical element of a language, such as a word, a phrase, or a prefix (see *Lexeme* on Wikipedia). Lexemes are Entities in the sense of the Wikibase data model. A Lexeme is described using the following information:

- An **ID**. Lexemes have IDs starting with an "L" followed by a natural number in decimal notation, e.g. `L3746552`. These IDs are unique within the repository that manages the Lexeme. The ID can be combined with a repository's concept base URI to form a unique URI for the Lexeme.

- A **Lemma** for use as a human readable representation of the lexeme, e.g. *"run"*.
- The **Language** to which the lexeme belongs. This is a reference to a concrete Item, e.g. Q1860 for *English*.
- The **Lexical category** to which the lexeme belongs. This is given as a reference to a concrete Item, e.g. Q34698 for *adjective*.
- A list of Statements to describe properties of the lexeme that are not specific to a Form or Sense (e.g. *derived from* or *grammatical gender* or *syntactic function*)
- A list of Forms, typically one for each relevant combination of grammatical features, such as *2nd person / singular / past tense*.
- A list of Senses, describing the different meanings of the lexeme (e.g. "financial institution" and "edge of a body of water" for the English noun *bank*).



visualization of the Lexeme data model

Editorial Note: We should provide some hint regarding how grammatical gender can be modeled using Statements.

Lemma

The lemma is a human readable representation of the lexeme (see *Lemma* on Wikipedia). Typically, the canonical form of the lexeme (e.g. the infinitive form of verbs) will be used as the lemma (see also lemon:canonicalForm (https://www.w3.org/2016/05/ontolex/#canonicalForm)). Lemmas are not simple strings, but MultilingualTextValues, since the same lemma may have multiple spellings. This is specially important for languages that use multiple scripts such as Serbian and Japanese.

Example: The Lemma for English noun *color* would include "colour" for British English as well as "color" for American English.

A Lemma cannot be entirely empty, at least one variant has to be provided.

Note: Lemmas are not unique, nor is the combination of Lemma, Language, and Lexical category. Two distinct lexemes with the same lexical category can exist in the same language if they have different data, it may be gender, etymology, morphology (different forms), and so on.

Example: There are two German nouns with the Lemma "See", differing only in gender: "der See" meaning "the lake", and "die See" meaning "the sea". These two meanings cannot be understood as a single Lexeme, since they have different forms according to their gender.

Form

The morphology of the lexeme is understood as a set of Forms. Each form defines how a lexeme changes based on a specific *syntactic role* or *mode* it may take in a sentence (see also lemon:Form (https://www.w3.org/2016/05/ontolex/#Form)).

Example: The English verb *run* becomes "running" as a *present participle* and "runs" in *3rd person singular*.

A Form is described using the following information:

- An **ID**. Forms have IDs starting with the ID of the Lexeme they belong to, followed by a hyphen ("-") and an "F", followed by a natural number in decimal notation: e.g. `L3746552-F7`. These IDs are unique within the repository that manages the Lexeme. The ID can be combined with a repository's concept base URI to form a unique URI for the Form.
- A **representation**, spelling out the Form as a string.
- A list of **grammatical features** that define for which syntactic role the given form applies. These are given as references to a concrete Items, e.g. Q814722 for *participle*.
- A list of Statements further describing the Form or its relations to other Forms or Items (e.g. *pronunciation audio*, *rhymes with*, *used until*, *used in region*)

Planned Feature:
We may add the notion of a "form type" that determine what information a Form contains. One possible new type could be "nonexistent", which would allow to to represent forms that are known to not exist (like the infinitive of English "may", or the plural of German "Schnee"). Forms of the "nonexistent" kinds would have statements and grammatical features, but no representations.

Representation

A form's Representation is its written form, as used in a text (compare lemon:writtenRep (https://www.w3.org/2016/05/ontolex/#writtenRep)). Just like Lemmas, Representations are not simple strings, but MultilingualTextValues, since the same form may have multiple spellings, possibly in multiple scripts.

A Representation cannot be entirely empty, at least one variant has to be provided.

Grammatical Feature

A form's grammatical features specify under which conditions or in which syntactic role that form is used (see lexinfo:morphosyntacticProperty (http://lexinfo.net/ontology/2.0/lexinfo.owl#morphosyntacticProperty) and grammatical category on Wikipedia). Multiple grammatical features can be combined to express under which conditions the language's grammar requires a given form to be used. Grammatical features are represented as references to Items.

Example: The role *1st person present tense plural* can be defined by three features, represented by Wikidata Items: Q192613 (present tense), Q21714344 (first person), and Q146786 (plural).

Editorial Note: How do we model "a" vs "an"? What item would we use as a feature to describe this? Do we need free text usage notes after all?

Editorial Note: We should note that gender-specific forms like "baroness" can be treated as Forms, or as separate Lexemes, as need be.

Sense

The senses of a lexeme are different meanings which it may represent in a text. The senses are given as natural language definitions or *glosses* (compare *intensional definitions* on Wikipedia).

A sense is described using the following information:

- An **ID**. Senses have IDs starting with the ID of the Lexeme they belong to, followed by a hyphen ("-") and an "S", followed by a natural number in decimal notation: e.g. `L3746552-S4`. These IDs are unique within the repository that manages the Lexeme. The ID can be combined with a repository's concept base URI to form a unique URI for the Sense.
- A **Gloss**, defining the meaning of the Sense using natural language.
- A list of Statements further describing the Sense and its relations to Senses and Items (e.g. *translation*, *synonym*, *antonym*, *connotation*, *register*, *denotes*, *evokes*).

Editorial Note: We should find a good place to address a common source of misunderstandings: Senses can be connected to Wikidata Items via an appropriate Statement they evoke or denote (compare lemon:denotes (https://www.w3.org/2016/05/ontolex/#denotes) and lemon:evokes (https://www.w3.org/2016/05/ontolex/#evokes)). However, such a connection should not be interpreted as the lexeme actually representing the concept defined by the item (compare lemon:LexicalSense (https://www.w3.org/2016/05/ontolex/#LexicalSense) and lemon:LexicalConcept (https://www.w3.org/2016/05/ontolex/#LexicalConcept)). In particular, if two lexemes have senses that refer to the same concept in this way, this does not imply that the two lexemes are synonyms.

Example: The lexemes for the English adjectives "hot" and "cold" could both have a sense that refers to Q11466 (temperature), even though they are antonyms.

Editorial Note: We should describe how word *function* can be described for things like "to" or "a", using Statements on the Lexeme. We should also explain that function words should not have senses. Do we need free text usage notes?

Planned Feature:
We may introduce a field in the Sense for syntactic markers and/or syntactic frames for subcategorization (see also the definition (http://www.unlweb.net/wiki/Subcategorization_frames) on the UNL wiki). That would allow "ask for", "ask about", "ask to", "ask out", "ask oneself", etc. to be modeled as sense of the same lexeme, each with a different subcategorization. Some verbs also change the meaning depending on whether they are used reflexively (e.g. German "übernehmen" vs "sich übernehmen"). Compare synsem:marker (https://www.w3.org/2016/05/ontolex/#marker) and synsem:syntactic-frame (https://www.w3.org/2016/05/ontolex/#syntactic-frames).

Gloss

A sense's gloss gives a natural definition of the sense (see *Gloss* on Wikipedia and skos:definition (https://www.w3.org/2009/08/skos-reference/skos.html#definition)). Similar to Lemmas, Glosses are not simple strings, but MultilingualTextValues. However, the reason is not providing support for variants, but to allow the gloss to be given in entirely different languages. E.g. it would be quite useful for a German learning French to have a German gloss for a French word.

A Gloss cannot be entirely empty, at least one language has to be provided.

See also

- Lexeme data model examples
