Wikibase/DataModel/Primer

< Wikibase | DataModel

> This document is a draft, and should not be assumed to represent the ultimate structure.

This is a primer to the Wikibase data model. For a more technical specification please check the data model specification.

Contents

Summary of the data model

Wikibase knowledge base content can be summarized as follows:

A Wikibase knowledge base is a collection of Entities. **Entities** are the basic elements of the knowledge base, which can be described and referenced using the Wikibase data model. There are two predefined kinds of Entities: Items and Properties. Wikibase may be extended to support additional types of Entities.

The description of Items and Properties are structured as follows.

1. **Item**:
    1. **Item identifier** (number prefixed with *Q*)
    2. **Fingerprint**, consisting of:
        1. Multilingual **label***
        2. Multilingual **description***
        3. Multilingual **aliases**
    3. **Statements**, each consisting of:
        1. **Claim**, consisting of:
            1. Property

           2. Value
           3. Qualifiers (additional property-value pairs)
       2. **References** (each consisting of one or more property-value pairs)
       3. **Rank**
   4. **Site links**
2. **Property**

   1. **Property identifier** (number prefixed with *P*)
   2. **Fingerprint**, consisting of:

      1. Multilingual **label***
      2. Multilingual **description***
      3. Multilingual **aliases**
   3. **Statements**, each consisting of:

      1. **Claim**, consisting of:

         1. Property
         2. Value
         3. Qualifiers (additional property-value pairs)
      2. **References** (each consisting of one or more property-value pairs)
      3. **Rank**
   4. **Datatype**

*) Unless label and/or description of an entity are not empty, within the scope of an entity type, an entity's combination of label and description in a certain language must be unique.

Items

One page in Wikibase describes one **item**. Items are the way Wikibase refers to anything of interest, and usually are the things that Wikipedia articles are about. So in Wikibase we will have an item for Berlin, and what we mean with this item is the topic of the Wikipedia articles linked to this item in the different languages. The Wikipedia articles identify the meaning of an item.

Every item has a **label** (a name) and a **description** in each language of Wikibase. Just the label would not be enough as it may be ambiguous: Berlin could refer to the capital of Germany, one of more than a dozen cities in the US, a Lou Reed album, an American new wave band, or many other things. The label and the description together should identify the meaning of an item, e.g. the label "Berlin" and the description "A city in Germany" should be uniquely identifying in each language.

In addition to labels, items can have **aliases** which provide alternative names for an item to be found. *"George H. W. Bush"* might also be found under *"George Bush"*, and so might his son. Aliases are meant to offer the user search convenience, much like redirects on Wikipedia, and thus even popular misspellings may be used as aliases.

The symbol grounding problem

If you are following carefully you will notice that both the Wikipedia links and label plus description identify the meaning of an item. And not only that: they do that in all languages! It can thus happen that these identifiers get out of sync: the German Wikipedia link might point to Berlin, Kentucky and the English description might say *"Capital of Germany".* This is true, and there is nothing implemented in the system to prevent it: no language and no identifying mechanism has precedence over the other. Here we are running into the symbol grounding problem. The path we are taking in Wikibase to address this problem is by deliberatively providing multiple ways to identify the meaning of an item and trust that Wikibase editors will come up with a socio-technical mechanism to solve it well enough for the Wikibase use cases.

Statements
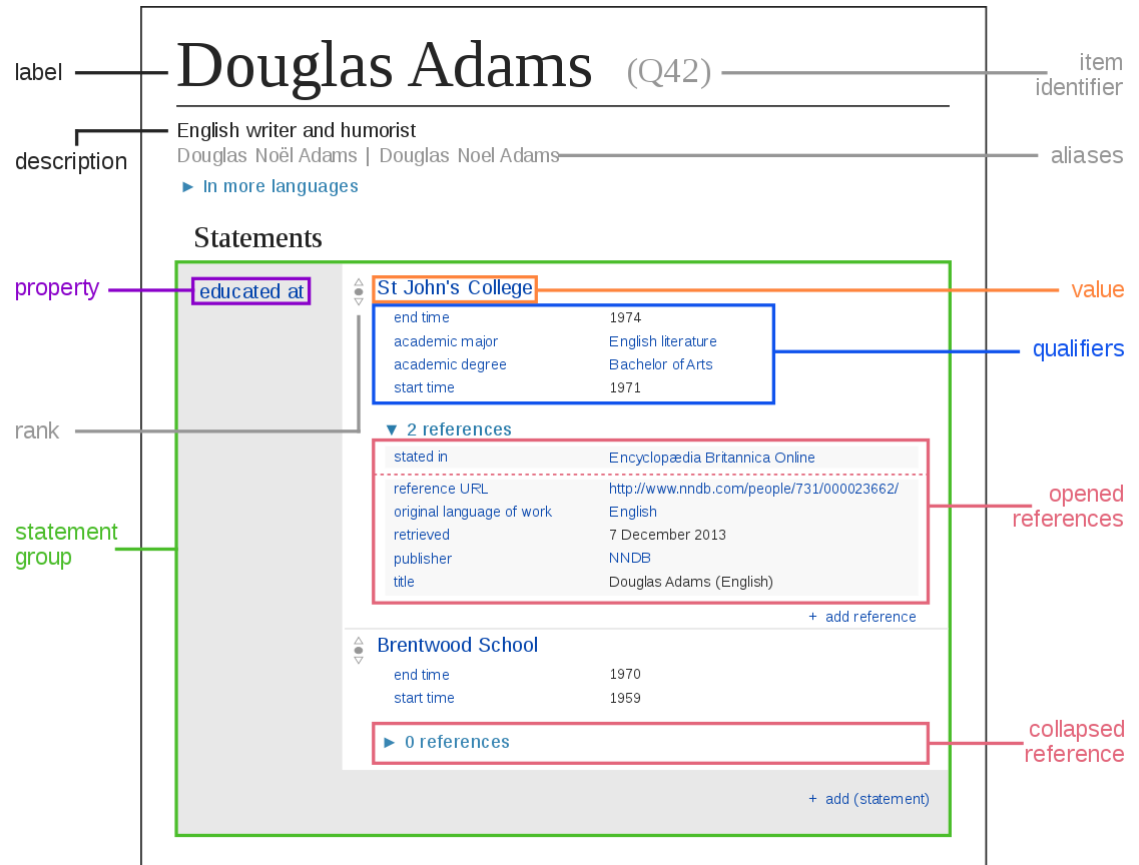
Overview of a Wikibase statement

One of the requirements is that *"Wikibase will not be about the truth, but about statements and their references."* This means that in Wikibase we do not actually model the *items themselves,* but *statements about them.* We do not say that Berlin has a population of 3,5 M, we say that there is this statement about Berlin's population being 3,5 M as of 2011 according to the German statistical office.

A **statement** may consist of

- one property (in the example, "population")
- one value (3,5 M)
- optionally one or more qualifiers (in this example, "as of 2011" is one of the qualifiers)
- optionally one or more references (the Germans statistical office)

The property, value, and qualifiers together are also called the **claim**, which together with any source references forms a statement.

There can be several statements about the same property: people can have several children, books might have several authors. Also, there might be diverging points of view on the population of a city -- official numbers and UN estimates, for example. Or there might be values with different qualifiers, like points in time or measurement methods. For a few examples, see below.

Properties are described on their own wiki pages in Wikibase. Properties also have labels and descriptions, and additionally to that they also have a data type associated with them and perhaps additional properties. The data type defines the type of the value used with this property. The set of properties is created and maintained by the Wikibase editors.

Values themselves can be either very simple -- another item or just a string -- or quite complex beasts, like a geographic shape, a measurement with a unit and an accuracy, or a time period. We will describe values in more detail in their own page in the future. The set of data types is (mostly) predefined.

There are two special values, mostly regardless of their data type: **none** and **unknown**. *None* means that we know that the given property has no value, e.g. Elizabeth I of England had no spouse. *Unknown* means that the property has a value, but it is unknown which one -- e.g. Pope Linus most certainly had a year of birth, but it is unknown to us. This should not be mixed up with the notion that it is unknown whether an item has a value for a specific property, e.g., if a person had children. Both *none* and *unknown* are also not to be confused with the respective string: having the name *"unknown"* is different from having an unknown name (which is again different from it being unknown whether the entity has a name).

References offer a source that supports the given claim. There can be several references given for a statement. We are still working on how to further structure a reference, but in general they will point to a source (which would be a Wikibase item in its own right, e.g. a book, a website, etc.) and have further information, like the page where the claim is supported, etc. A claim without references is not necessarily wrong, nor is a claim with references true. It is still up to the reader of the statement to decide if they want to trust the claim or not. We will describe references in more detail in their own page in the future.

Example statements

Two statements without qualifiers:

Berlin

Area	891.85 km²	[1 source]
Mayor	Michael Müller	[no sources]

One statement with two qualifiers:

Germany

Chancellor	Angela Merkel	[2 sources]
	since 2005	
	Party CDU	

Two statements with the same property, each with one qualifier:

Berlin

Population	3,500,000	[no sources]
	as of 2012	
	8,000	[1 source]
	as of 15th century	

Qualifiers

Qualifiers are used to further describe or refine the value of a property given in a statement. They consist of a property and a value, which are the same as for statements.

While it would be convenient if we could express all the data we need for the use cases of Wikibase with simple property-value pairs, this is unfortunately not the case. Many statements require further qualifiers in order to be expressed. In order to reduce the number of properties to a manageable size, qualifiers are used to further specify the statement in some way. Qualifiers can be used in a number of ways, as shown by the following examples.

A qualifier can modify what the item means (*"France: Area 213,010 sq mi - excluding Adélie Land"*), the property (*"Berlin: Population 3,500,000 - method Estimation"*), constrain the validity of the value (*"Germany: Population 80,000,000 - as of 2011"*), or offer further details (*"Austria: Religion Catholic - Percentage 64,8%"* or *"Goldfinger: Actor Sean Connery - Role James Bond"*), etc. A catch-all qualifier is expected to be "annotation" or something similar.

It is open to the Wikibase community to maintain and use qualifiers in a way that makes sense to them and for their use cases. The qualifier is an integral part of the statement: take away the qualifier, and the meaning of the statement is changed. This is far less true for the references.

Ranks

As there are potentially many different statements for a given item and property, we need to select which ones to return when Wikibase gets asked. In order to facilitate this, three **ranks** of statements are introduced. There can be any number of statements in each rank, but within each rank, their order is not significant.

- **Preferred statements**: if preferred statements exist, these statements are returned in response to a query. They would, e.g. for a population contain the most recent one as long as it is regarded as sufficiently reliable. Wikibase editors might decide to mark several statements as preferred: this may be used to indicate disagreement, reflecting the knowledge diversity on the issue, or it may be used to express the notion of actually having multiple values (in case of properties like "children").
- **Normal statements**: if there are no preferred statements (or the query explicitly says to include normal statements too), these statements are returned. Historical values, like the population of a country in the past, might be here, as well as less representative sources which are still considered relevant.
- **Deprecated statements**: for statements that are being discussed, or known to be erroneous, but still listed for the sake of completion or in order to prevent them being constantly added and removed. Deprecated statements only appear in search results if they are explicitly added or if they are selected based on their source. A footnote qualifier should usually accompany other-ranked statements.

Within Wikibase, the ranks are also used to make the display cleaner. Only the preferred statements are displayed by default, and the reader has to click on a link like *"more values"* in order to see the normal-ranked statements.

See also