Data modelling

Camillo Carlo Pellizzari di San Girolamo (user:Epìdosis)
Scuola Normale Superiore

Summary

1. Different levels: Wikibase, general and thematic properties
2. Wikibase: items, properties, ranks, qualifiers, references
3. General properties: “instance of” and “subclass of”
4. Thematic properties: constraints and WikiProjects
5. Notability criteria

1) Different levels

Wikibase, general and thematic properties

Premise

When modelling data on Wikidata, you have to consider:

- what is doable from a **technical perspective** (i.e. what the software used by Wikidata, Wikibase, allows to do)
- what is doable from a **community perspective** (i.e. how the community has chosen certain ways and not others to model a specific information)

For general reference, https://www.wikidata.org/wiki/Wikidata:Data_model

Levels

More specifically, we should consider:

1. how data are modelled in Wikibase
2. how general data are modelled in Wikidata
3. how topic-specific data are modelled in Wikidata

The first two aspects are quite stable and fixed, whilst data models on specific themes can often be improved through community discussions.

2) Wikibase

items, properties, ranks, qualifiers, references

Wikibase

Wikibase is the software used by Wikidata

- there are 3 types of entities: **items** (Q), **properties** (P), **lexemes** (L) – in this presentation lexemes will not be treated
- items (and lexemes) are used to store structured data, mainly through the use of properties

For technical reference,
https://www.mediawiki.org/wiki/Wikibase/DataModel

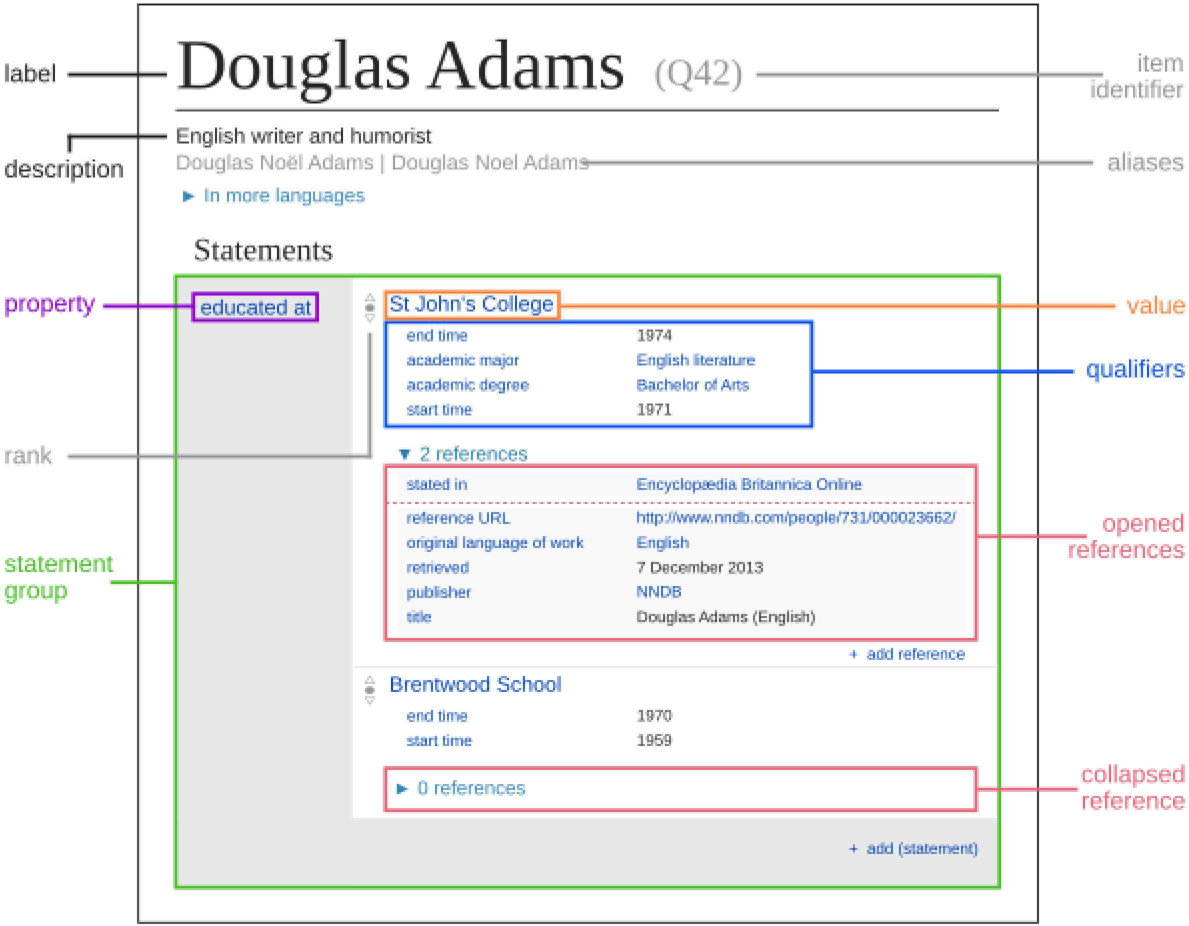
Items

Items are identified by **Q** followed by a number (QID).

Each item contains:

- labels, descriptions and aliases in different languages
- statements (divided into “Statements” and “Identifiers”)
- sitelinks

Statements have ranks and can have qualifiers and references



label
Douglas Adams
(Q42)
item identifier
description
English writer and humorist
Douglas Noël Adams | Douglas Noel Adams
aliases
In more languages
Statements
property
educated at
St John's College
value
end time 1974
academic major English literature
academic degree Bachelor of Arts
start time 1971
qualifiers
rank
2 references
stated in Encyclopædia Britannica Online
reference URL http://www.nndb.com/people/731/000023662/
original language of work English
retrieved 7 December 2013
publisher NNDB
title Douglas Adams (English)
opened references
statement group
+ add reference
Brentwood School
end time 1970
start time 1959
0 references
collapsed reference
+ add (statement)

A few definitions

- **Label**: preferred name of the entity in language X (cf. MARC21 for Authority Data field 100)
- **Description**: brief description of the entity in language X, mainly for disambiguation purpose
- **Aliases** (Also knowns as): alternative names of the entity in language X (cf. MARC21 for Authority Data field 400)

A few definitions

- **Statement**: a structured information about the entity; it consists mainly of a claim, which is a couple property-value (the value can be another item, a string, a date, a URL etc.)
- **Rank**: a way to annotate multiple values of a statement
- **Qualifier**: the contextualization of a statement, expressed through a couple property-value
- **Reference**: the source of a statement, expressed through one or more couples property-value

Properties

Properties are identified by **P** followed by a number.

Each property contains:

- labels, descriptions and aliases in different languages
- a datatype
- statements (divided into “Statements” and “Constraints”)

Datatypes

The datatype of a property defines what can be used as value of that property. The most used datatypes are:

- external ID
- item

Other datatypes include: quantity, string, URL, Commons file, point in time, monolingual text, coordinates.

For general reference, https://www.wikidata.org/wiki/Help:Data_type

3) General properties

“instance of” and “subclass of”

Classification of items

In order to give a basic understanding of what an item is, two fundamental properties are used (and often confused):

- instance of (P31)
- subclass of (P279)

A fundamental guideline explains how they should be used:
https://www.wikidata.org/wiki/Help:Basic_membership_properties

«Instance of» and «subclass of»: definitions

- “instance of” (P31) means that the item is an individual belonging to a certain class

e.g. “Mahatma Gandhi” (Q1001) is instance of “human” (Q5)

- “subclass of” (P279) means that all the instances of this item belong to a certain class

e.g. “submarine volcano” (Q674775) is subclass of “volcano” (Q8072)

«Instance of» and «subclass of»: difference

- an item can have both “instance of” and “subclass of”

e.g. “biologist” (Q864503) is instance of “profession” (Q28640)
is subclass of “scientist” (Q901)

- an item representing an individual entity (e.g. a specific human, organization, place etc.) <u>only</u> has “instance of”

4) Thematic properties

constraints and WikiProjects

Premise

- A significant part of the usefulness of structured data lies in their **internal coherence**: the same information should not be structured in many different ways, otherwise it becomes very difficult to find and query all its occurrences
- However, Wikibase is designed to be very flexible: the online technical restriction to the use of properties is their datatype (which is a very loose restriction)

One abstract example

Abstractly, all of these could be valid statements:

- “instance of” (P31) “mayor” (Q30185)
- “occupation” (P106) “mayor” (Q30185)
- “position held” (P39) “mayor” (Q30185)

However, it is clear that allowing more than one possibility would be a problem for the quality of Wikidata

Regulating the use of thematic properties

In order to ensure that properties are used coherently, Wikidata has introduced (in 2017) the **property constraints**.

- Constraints are statements in property pages; they specify how the properties should generally be used, restricting their abstractly possible values in many different ways.

For general reference,
https://www.wikidata.org/wiki/Help:Property_constraints_portal

Solution of the previous example

So, for “mayor” (Q30185), we can be guided by constraints:

- “instance of” (P31) should not be used: in items about humans, the only value of P31 should be “human” (Q5)
- “occupation” (P106) should not be used: the value of P106 should be an occupation (as politician), not a position (as mayor)
- “position held” (P39) should be used, since mayor is a position

Discovering thematic properties

In order to know which properties are used to describe a certain type of item, you can try to find a related WikiProject.

- **WikiProjects** are pages used by groups of users interested in the same topic to collect documentation about how that topic is presently treated in Wikidata and to discuss related problems and possible solutions.

WikiProjects

WikiProjects often include lists of properties that should be used to describe the related types of items; see e.g.

- https://www.wikidata.org/wiki/Wikidata:WikiProject_Books
- https://www.wikidata.org/wiki/Wikidata:WikiProject_Periodicals
- https://www.wikidata.org/wiki/Wikidata:WikiProject_Built_heritage

You can navigate WikiProjects through https://www.wikidata.org/wiki/Wikidata:WikiProjects

Other ways to discover thematic properties

You can discover which properties should be used in a certain type of item also:

- looking at **similar items** and imitating them (making sure that the imitated item has no constraint-violation icons)
- or just adding a new statement and selecting the **property suggested** by Wikidata itself, if it makes sense

If you add a constraint violation, don't worry and try to understand how to fix it looking at the message generated by the icon.

5) Notability criteria

Which items are accepted on Wikidata

https://www.wikidata.org/wiki/Wikidata:Notability

The notability criteria (as of September 2024)

An item is acceptable if and only if it meets **at least one** of the three criteria below:

1. it contains at least **one valid sitelink** to a page of a Wikimedia project (Wikipedia etc.);
2. it refers to an instance of a **clearly identifiable conceptual or material entity** that can be described using **serious and publicly available references**;
3. it fulfills a **structural need**.

Thanks for your attention!

Camillo Carlo Pellizzari di San Girolamo
camillo.pellizzaridisangirolamo@sns.it