Data modeling in Wikidata

Requirements for a Wikidata schema language

Jakob Voß

Verbundzentrale des GBV – VZG

Workshop on data quality management in Wikidata

2019-01-18, Berlin

Data modeling in Wikibase instances: Proposal of a Wikibase database language

Scheme language for *any* Wikibase instance

$\longrightarrow$ *Database language* for Wikibase

$\Longrightarrow$ Quality control in Wikidata

Wikibase is a database management system, so…

- …what’s it’s *database language*?!
- …where’s the *command line*?!

Database languages

- ~~**DCL** Data control language (access)~~
- **DDL** Data definition language (schema language)
- **DML** Data manipulation language (editing)
- **DQL** Data query language

Application in Wikidata

1. Data manipulation (editing)
2. Data query (SPARQL…)
3. Data definition (consistency, constraints…)

Database languages for Wikibase

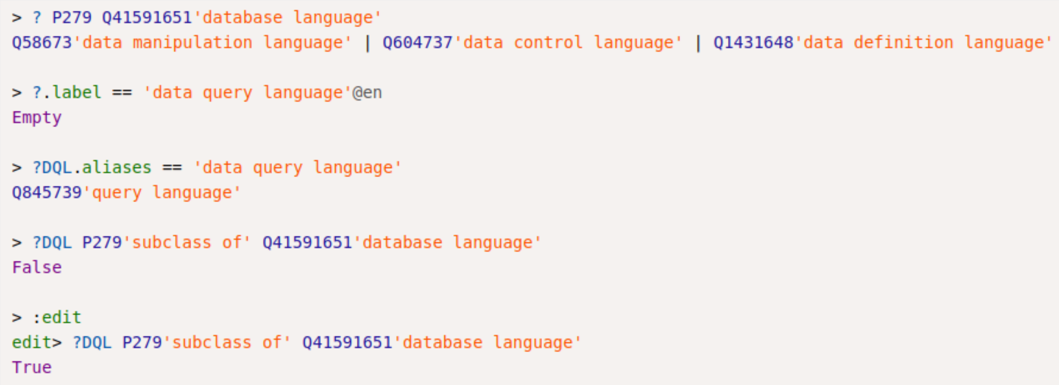
- **manipulation languages**
  QuickStatements, wikidata-cli, scripts…
- **query languages**
  SPARQL, GraphQL, wikidata-cli, scripts…
- **schema languages**
  property statements & constraints, SheX, scripts…

scripts: JavaScript, Java, Python, Lua, .NET…

Why another database language for Wikibase?

- Unified syntax for querying, editing & rules
- Current languages are bound to another level of abstraction
  - serialization formats (JSON, RDF)
  - programming languages

How about something like this?



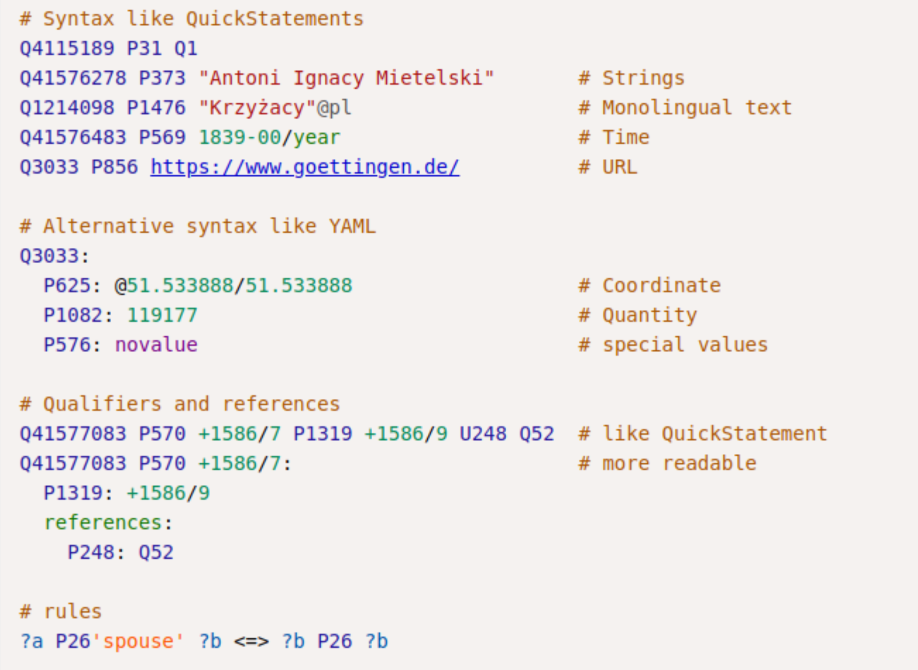
```
> ? P279 Q41591651'database language'
Q58673'data manipulation language' | Q604737'data control language' | Q1431648'data definition language'

> ?.label == 'data query language'@en
Empty

> ?DQL.aliases == 'data query language'
Q845739'query language'

> ?DQL P279'subclass of' Q41591651'database language'
False

> :edit
edit> ?DQL P279'subclass of' Q41591651'database language'
True
```

What’s wrong with existing languages?

- If you have a SPARQL & SheX, everything looks like RDF
- If you have a JavaScript, everything looks like JSON

Confusion of abstraction

- The Wikibase data model is not RDF, JSON, SQL, CSV… but a data model of its own
- Wikibase data language should build on the Wikibase data model
- Syntax matters!



```
# Syntax like QuickStatements
Q4115189 P31 Q1
Q41576278 P373 "Antoni Ignacy Mietelski"       # Strings
Q1214098 P1476 "Krzyżacy"@pl                   # Monolingual text
Q41576483 P569 1839-00/year                    # Time
Q3033 P856 https://www.goettingen.de/          # URL

# Alternative syntax like YAML
Q3033:
  P625: @51.533888/51.533888                   # Coordinate
  P1082: 119177                                # Quantity
  P576: novalue                                # special values

# Qualifiers and references
Q41577083 P570 +1586/7 P1319 +1586/9 U248 Q52  # like QuickStatement
Q41577083 P570 +1586/7:                        # more readable
  P1319: +1586/9
  references:
    P248: Q52

# rules
?a P26'spouse' ?b <=> ?b P26 ?b
```

Features of a Wikibase database language

- Wikibase data types as core datatypes
  - entities (items, properties, lexemes…)
  - string, monolingual, quantity (⊃ numbers!)
  - coordinates, times
  - media, tabular, shape…
- labels, aliases, claims, sitelinks…
- ranks, novalue, somevalue…
- rules to express expectations
- concise and readable syntax

Current state

- Draft of a specification
  - https://github.com/wikicite/kukulu/ (sources)
  - http://wikicite.org/kukulu/ (HTML)
- A buggy parser and syntax highlighter
  - written in NodeJS with chevrotain
  - throw-away prototype for experimenting
- An available name and its abbreviation (`kkl`):

kūkulu (Hawaiian): to build, to construct