Database Validation: Interesting facet of counter-vandalism

Case Study of Wikidata

Houcemeddine Turki, University of Sfax, Tunisia
Database Validation

A process to ensure data quality of a given semantic resource by managing competency questions

- **Accuracy**: Verification of whether the definitions, classes, properties and individual entries in the assessed resource are correct or not.
- **Completeness**: Coverage of a given knowledge domain in the evaluated resource.
- **Adaptability**: Range of different anticipated uses of the evaluated resource.
- **Clarity**: Effectiveness of communication of intended meanings of defined items by the assessed resource.
Main Purpose

Data Quality

- Ensuring an homogenous and exhaustive representation of structured information
- Eliminating redundancies and logical inconsistencies
- Verifying the accuracy of structured data
- Enhancing the reliability and trustworthiness of structured information

Counter-vandalism

- Ensuring that users do not change more accurate information by less accurate data
- Can also be evaluated by analyzing user behaviors and patrolling edits from non confirmed editors: [https://www.wikidata.org/wiki/Wikidata:Patrol](https://www.wikidata.org/wiki/Wikidata:Patrol)
- Further information can be found at [https://www.wikidata.org/wiki/Wikidata:WikiProject_Counter-Vandalism](https://www.wikidata.org/wiki/Wikidata:WikiProject_Counter-Vandalism)
Wikidata is linked to numerous external resources: Wikidata statements can be verified against information in external databases.
**Property statements and constraints**

A set of semantic information defining the format of the statements using a given Wikidata property

<table>
<thead>
<tr>
<th>Property Type</th>
<th>Example</th>
<th>Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>subject item of this property</td>
<td>affiliation</td>
<td>0 references</td>
</tr>
<tr>
<td>Wikidata property example</td>
<td>Raoul Bost</td>
<td>Institute for Advanced Study</td>
</tr>
<tr>
<td>subproperty of</td>
<td>use</td>
<td>0 references</td>
</tr>
<tr>
<td>Wikidata property example</td>
<td>dabrafenib</td>
<td>skin melanoma</td>
</tr>
<tr>
<td>equivalent property</td>
<td>0 references</td>
<td></td>
</tr>
<tr>
<td>inverse property</td>
<td><a href="http://purl.obolibrary.org/obo/RO_0002599">http://purl.obolibrary.org/obo/RO_0002599</a></td>
<td>0 references</td>
</tr>
<tr>
<td>value type constraint</td>
<td>organization</td>
<td></td>
</tr>
<tr>
<td>class</td>
<td>fictional organization</td>
<td></td>
</tr>
<tr>
<td>project</td>
<td>fictional municipal police</td>
<td></td>
</tr>
<tr>
<td>fictional gang</td>
<td></td>
<td></td>
</tr>
<tr>
<td>institution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>institute</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Christian ministry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>church building</td>
<td></td>
<td></td>
</tr>
<tr>
<td>place of worship</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fictional place of worship</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Christian movement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>religious denomination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>instance of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>relation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>type constraint</td>
<td>human</td>
<td></td>
</tr>
<tr>
<td>class</td>
<td>group of humans</td>
<td></td>
</tr>
<tr>
<td>fictional organization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fictional character</td>
<td></td>
<td></td>
</tr>
<tr>
<td>organization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>instance of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>relation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A notification appears to Wikidata users when a given property constraint is violated.
Recoin

A tool to identify missing statements for a given item by comparing it to its class members

• Identifies the Wikidata items that are not significantly described and that are consequently likely to be created due to vandalism.
• Can help identifying properties that are not commonly used for the members of a given class.

2020 COVID-19 pandemic in Tunisia (Q87343682)
Shape Expressions (ShEx)
Structural schema language to define the format of the members of a given Wikidata class

```xml
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>

start = @<app>

<app> EXTRA wdt:P31 { 
  wdt:P31 [ wd:Q90790055 # instance of COVID-19 dashboard or 
    wd:Q91136116 # search engine or 
    wd:Q91137337 # dataset 
  ];
  wdt:P1476 LITERAL * ; #title 
  wdt:P366 . + ; #use 
  wdt:P123 . + ; #publisher 
  wdt:P178 . + ; #developers 
  wdt:P495 . + ; #country of origin 
  wdt:P306 . + ; #operating system 
  wdt:P856 . + ; #official website 
  wdt:P921 . + ; #main subject 
  wdt:P144 . + ; #based on 
  wdt:P577 . ? ; #publication date 
  wdt:P7103 . ? ; #start of covered period 
  wdt:P275 . + ; #copyright license 
  wdt:P5008 . + ; #on focus list of Wikimedia project 
}
```

<table>
<thead>
<tr>
<th>Key</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>wdt:&lt;PropertyID&gt;</td>
<td>Defined property</td>
</tr>
<tr>
<td>wd:&lt;ItemID&gt;</td>
<td>Defined item</td>
</tr>
<tr>
<td>.</td>
<td>Object</td>
</tr>
<tr>
<td>*</td>
<td>Zero or more</td>
</tr>
<tr>
<td>?</td>
<td>Zero or one</td>
</tr>
<tr>
<td>+</td>
<td>One or more</td>
</tr>
<tr>
<td>LITERAL</td>
<td>Monolingual text</td>
</tr>
<tr>
<td>EXTRA</td>
<td>Object is one value from of a defined list</td>
</tr>
</tbody>
</table>

Shape Expressions (ShEx)

Structural schema language to define the format of the members of a given Wikidata class

**Shape Expression for class**  [edit]

Originally proposed at Wikidata:Property proposal/Generic

<table>
<thead>
<tr>
<th>On hold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
</tr>
<tr>
<td>Represents</td>
</tr>
<tr>
<td>Data type</td>
</tr>
<tr>
<td>Domain</td>
</tr>
<tr>
<td>Example 1</td>
</tr>
<tr>
<td>Example 2</td>
</tr>
<tr>
<td>Example 3</td>
</tr>
<tr>
<td>Example 4</td>
</tr>
</tbody>
</table>

**Motivation**  [edit]

Property to link a class to the Shape Expression that members of it should conform to.

This will make it easier to query for Shape Expressions that exist, and quickly see what has been defined for a particular class. Jheald (talk) 16:56, 28 May 2019 (UTC)

Note: Implementation will require EntitySchemas to be added to the set of data-types that can be values for Wikidata statements. There is a ticket for this on Phabricator, which Léa hopes should be resolved in the coming weeks.[1][2]. Jheald (talk) 07:54, 29 May 2019 (UTC)

- A property to link EntitySchemas to Wikidata classes is currently on hold
- There is no need for this property if a script can be built to automatically validate Wikidata items against concerned EntitySchemas
Shape Expressions (ShEx)

Structural schema language to define the format of the members of a given Wikidata class

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>

start = @<app>

<app> EXTRA wdt:P31 {  
    wdt:P31 [ wd:Q90790055 # instance of COVID-19 dashboard or  
        wd:Q91136116 # search engine or  
        wd:Q91137337 # dataset  
    ] ;  
    wdt:P1476 LITERAL * ; #title  
    wdt:P366 . * ; #use  
    wdt:P123 . * ; #publisher  
    wdt:P178 . * ; #developers  
    wdt:P495 . * ; #country of origin  
    wdt:P306 . * ; #operating system  
    wdt:P856 . * ; #official website  
    wdt:P921 . * ; #main subject  
    wdt:P144 . * ; #based on  
    wdt:P577 . ? ; #publication date  
    wdt:P7103 . ? ; #start of covered period  
    wdt:P275 . * ; #copyright license  
    wdt:P5008 . * ; #on focus list of Wikimedia project  
}
```

**Interesting hint:** ShEx statements where the object is a defined Wikidata item (Not ?, *, or +) can be used to define the condition of the application of the EntitySchema.

Consistency rules

Conditions allowing the identification of data inconsistencies through the comparison of Wikidata statements

Example 1: For a given disease, the number of cases (P1603) in day Z should be inferior or equal to the one in day Z+1.

Example 2: For a given disease, the number of cases (P1603) in a given continent for day Z should be equal to the sum of the number of cases in the sovereign states in that continent in the same day Z.

Example 3: For a given disease, the number of deaths (P1120) and the number of recoveries (P8010) in a given location in day Z should be inferior or equal to the number of cases (P1603) in that location in the same day Z.

Example 4: If X is an instance of disease (P31 Q12136) and Y is the drug used for treatment (P2176) of X, Y should be an instance of a drug (P31 Q12140).

Example 5: If X is a drug that has dyspnea (Q188008) as a side effect (P1909), X cannot be the drug used for treatment (P2176) of asthma (Q35869) as well as of COPD (Q199804).

➔ Cannot be verified using Property Constraints or EntitySchemas
➔ Can be significantly implemented using SPARQL
A framework where consistency rules, property constraints and RDF validation schemas interact to validate semantic information will enhance Wikidata data quality.
Reference bots

A bot that can add references to Wikidata statements from scholarly databases

- When such algorithms cannot find a reference for a given statement, this statement is likely to be wrong
- When such algorithms find a reference for a given statement, this statement is likely to be correct
Bibliometric-Enhanced Validation of Wikidata statements

- Citation-Based Validation of main subject (P921) statements:
  - When a paper does not cite (P2860) a work linked to the same topic, the statement is likely to be wrong.
  - When a paper is not co-cited with another paper linked to the same topic, the statement is likely to be wrong.
  - When a paper is not cited by a work linked to the same topic, the statement is likely to be wrong.

- Semantic-Based Validation of main subject (P921) statements:
  - Sentence-level semantic similarity measures compute the level of similarity between sentences based on the characteristics of an is-a taxonomy.
  - ‘Subclass of (P279)’, ‘Part of (P361)’ or ‘Instance of (P31)’ Wikidata taxonomy can be used as a reference resource for this computation.
  - If the semantic similarity between the titles of two research papers represented in Wikidata is limited, the two scholarly publications are not likely to evoke the same research topic.
Thank You

Houcemeddine Turki
turkiabdelwaheb@hotmail.fr
+21629499418