Taxonomy browsing and ontology evaluation for Wikidata

Serghei Stratan

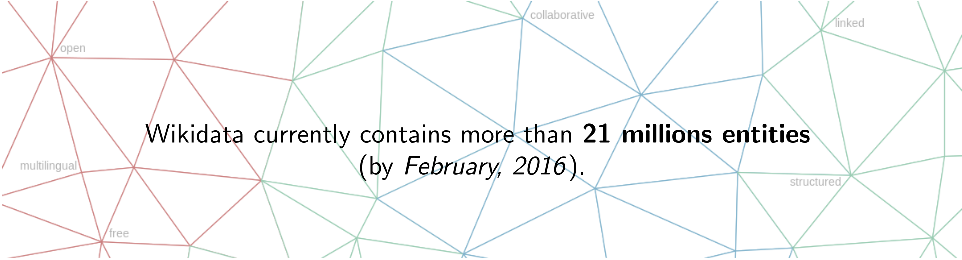
Technische Universität Dresden

February 12, 2016

Motivation

- How we can **browse the class hierarchy** from Wikidata?
- How can we **improve the quality of ontology** modelling in Wikidata?

Wikidata



Wikidata currently contains more than **21 millions entities** (by *February, 2016*).

- **Free** linked database
- **Collecting structured data**
- **Collaborative**
- **Multilingual**

The Wikidata Entities

a Wikidata page $\Longrightarrow$ an entity

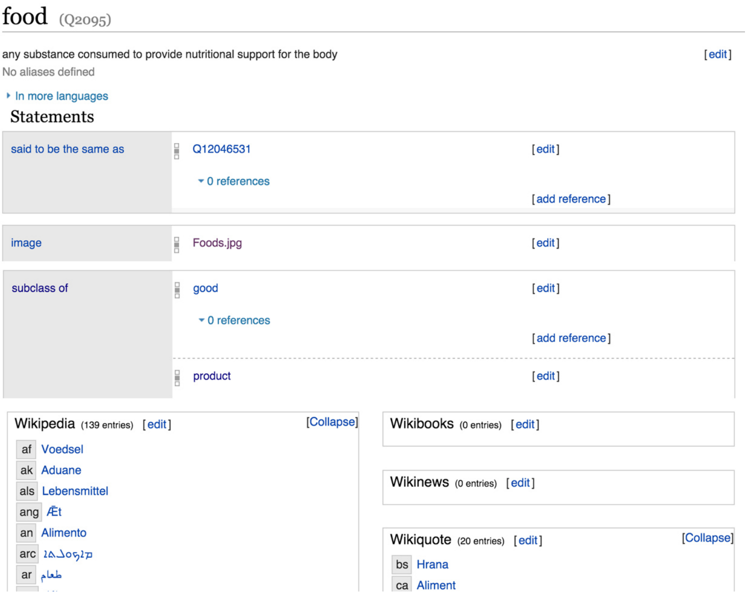
The Wikidata **items**: individuals and classes

- **Unique** – identifiable by a unique ID (with a **Q** prefix)
- **Notable –** usually have a corresponding page to some of the Wikimedia sites (Wikipedia, Wikivoyage, Wikisource etc.)
- **Linked**

The Wikidata **properties**: RDF properties

- Identifiable by a unique ID (with a **P** prefix)
- Have data types that determine the accepted value (string, URL, time, geographic coordinates etc.)

The Wikidata Item: food (Q2095)



food (Q2095)

any substance consumed to provide nutritional support for the body [edit]

No aliases defined

▸ In more languages

Statements

said to be the same as	Q12046531 ▾ 0 references	[edit] [add reference]
image	Foods.jpg	[edit]
subclass of	good ▾ 0 references	[edit] [add reference]
	product	[edit]

Wikipedia (139 entries) [edit] [Collapse]

af	Voedsel
ak	Aduane
als	Lebensmittel
ang	Ǣt
an	Alimento
arc	ܡܐܟܘܠܬܐ
ar	طعام

Wikibooks (0 entries) [edit]

Wikinews (0 entries) [edit]

Wikiquote (20 entries) [edit] [Collapse]

bs	Hrana
ca	Aliment

The Wikidata Class Hierarchy

Entities used:

- **Classes**: type of items which refers to a group of instances
- **Individuals**: individual instances or things or objects

Properties used:

- **subclass of (P279)**: similar to RDF `rdfs:subClassOf`
- **instance of (P31)**: similar to RDF `rdf:type`

Ontology Evaluation

Criteria for the Wikidata ontology evaluation:

1

2

3

4

5

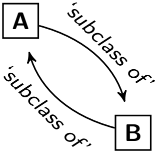
6

Ontology Evaluation

Criteria for the Wikidata ontology evaluation:

1. Cycles detection
2.
3.
4.
5.
6.

Cycle:

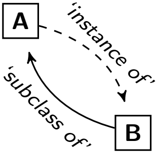


Ontology Evaluation

Criteria for the Wikidata ontology evaluation:

1. Cycles detection
2.
3.
4.
5.
6.

Cycle:

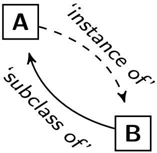


Ontology Evaluation

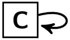
Criteria for the Wikidata ontology evaluation:

1. Cycles detection
2. Self-loops detection
3.
4.
5.
6.

Cycle:



Self-loop:

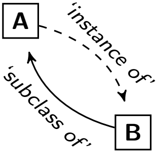


Ontology Evaluation

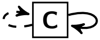
Criteria for the Wikidata ontology evaluation:

1. Cycles detection
2. Self-loops detection
3.
4.
5.
6.

Cycle:



Self-loop:

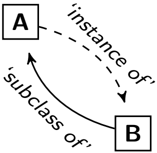


Ontology Evaluation

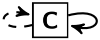
Criteria for the Wikidata ontology evaluation:

1. Cycles detection
2. Self-loops detection
3. Root classes

Cycle:



Self-loop:



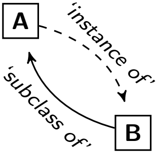
Root classes:

Ontology Evaluation

Criteria for the Wikidata ontology evaluation:

1. Cycles detection
2. Self-loops detection
3. Root classes
4. Finding classes which have more than 100 direct subclasses
5.
6.

Cycle:



Self-loop:

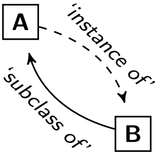
Root classes:

Ontology Evaluation

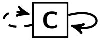
Criteria for the Wikidata ontology evaluation:

1. Cycles detection
2. Self-loops detection
3. Root classes
4. Finding classes which have more than 100 direct subclasses
5. Errors of relation properties

Cycle:



Self-loop:



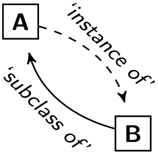
Root classes:

Ontology Evaluation

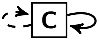
Criteria for the Wikidata ontology evaluation:

1. Cycles detection
2. Self-loops detection
3. Root classes
4. Finding classes which have more than 100 direct subclasses
5. Errors of relation properties
6. Redundancies of relation properties

Cycle:

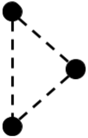


Self-loop:

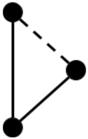


Root classes:

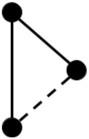
5. Error Patterns of relation properties



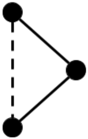
e1



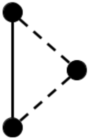
e2



e3

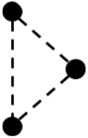


e4

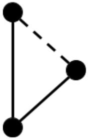


e5

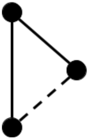
5. Error Patterns of relation properties



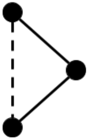
e1



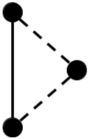
e2



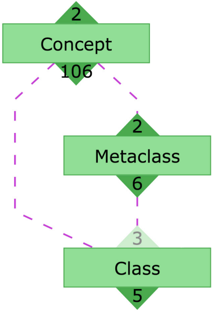
e3



e4

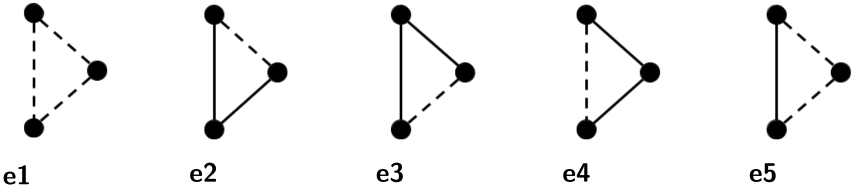


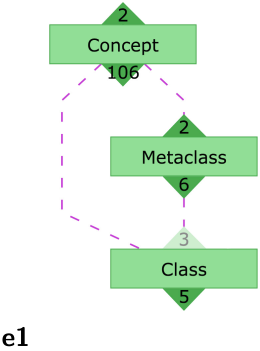
e5

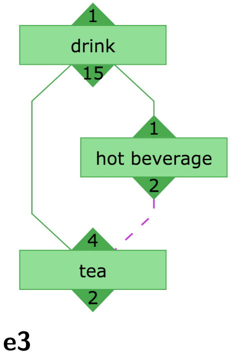


e1

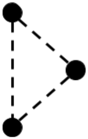
5. Error Patterns of relation properties



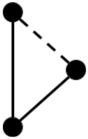




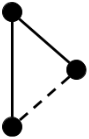
5. Error Patterns of relation properties



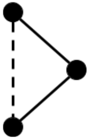
e1



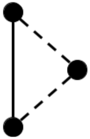
e2



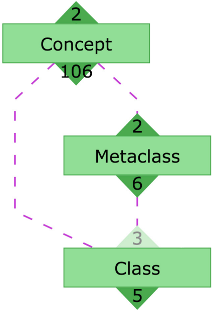
e3



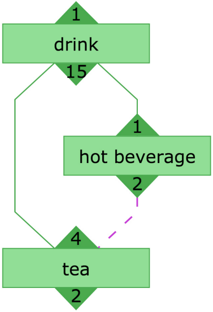
e4



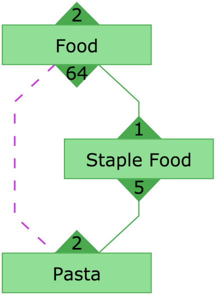
e5



e1

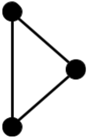


e3

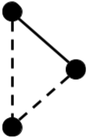


e4

6. Redundancy Patterns of relation properties

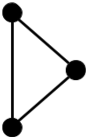


r1

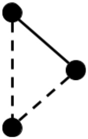


r2

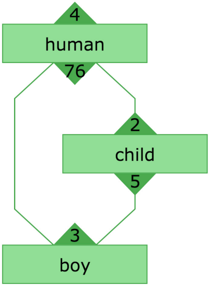
6. Redundancy Patterns of relation properties



r1

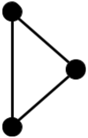


r2

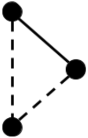


r1

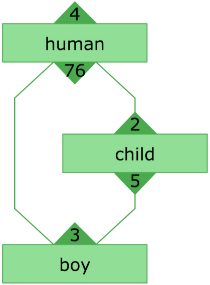
6. Redundancy Patterns of relation properties



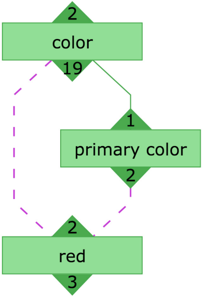
r1



r2

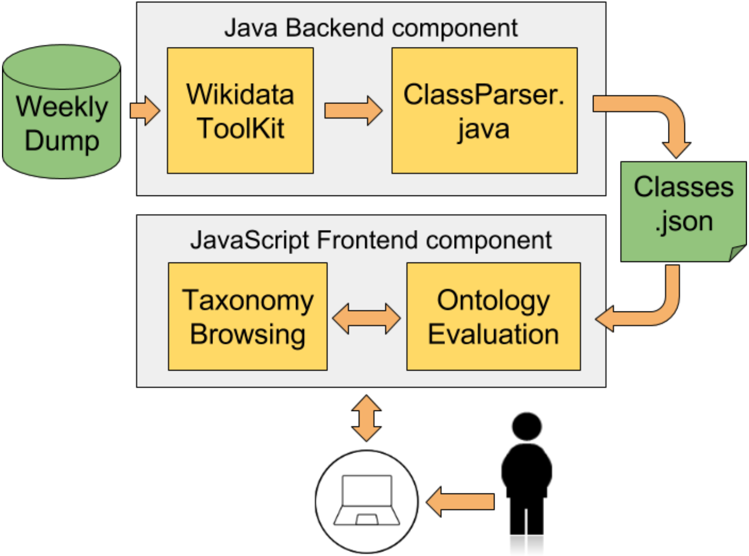


r1



r2

Architecture of the Developed System



Libraries and Tools

JavaScript libraries used in the *Frontend component*:

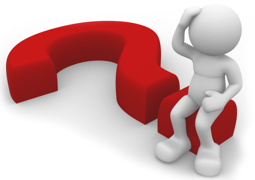
- **Dagre.js (version 0.0.6)** – lay out directed graphs
- **D3.js (version 3.0)** – graph visualization
- **jQuery (version 1.9.0)** – for general implementation
- **jQuery UI (version 1.11)** – user interface

Java libraries/tools used in the *Backend component*:

- **Java SE Development Kit (version 8u66)** – for developing
- **Wikidata Toolkit (version 0.4.0)** – access to the Wikidata repository and ontology dump files parsing

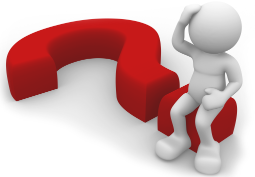
Technical Problems and Solutions

- Problem: *very slow response time* for data processing



Technical Problems and Solutions

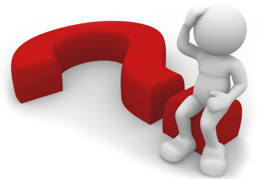
- Problem: *very slow response time* for data processing
- Solution: modified the JavaScript libraries (**Dagre.js** and **D3.js**)



Technical Problems and Solutions

- Problem: *very slow response time* for data processing
- Solution: modified the JavaScript libraries (**Dagre.js** and **D3.js**)

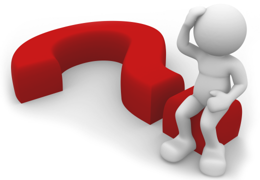
- Problem: *running out of virtual memory* for data extraction



Technical Problems and Solutions

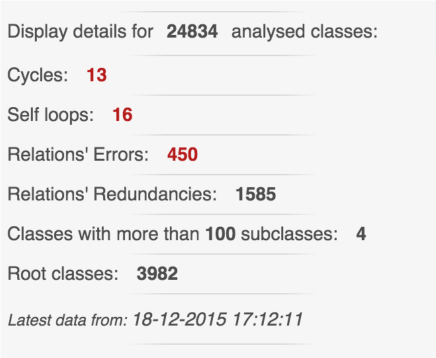
- Problem: *very slow response time* for data processing
- Solution: modified the JavaScript libraries (**Dagre.js** and **D3.js**)

- Problem: *running out of virtual memory* for data extraction
- Solution: convert extracted data fields



What we have learned about the Wikidata ontology

We found:



Display details for **24834** analysed classes:

Cycles: **13**

Self loops: **16**

Relations' Errors: **450**

Relations' Redundancies: **1585**

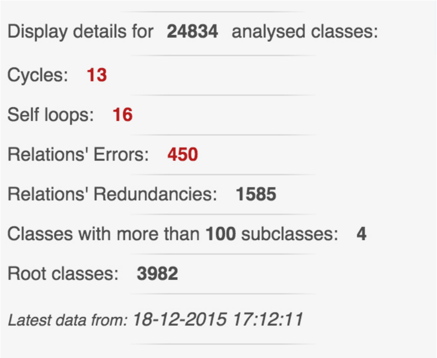
Classes with more than **100** subclasses: **4**

Root classes: **3982**

Latest data from: 18-12-2015 17:12:11

What we have learned about the Wikidata ontology

We found:



Display details for **24834** analysed classes:

Cycles: **13**

Self loops: **16**

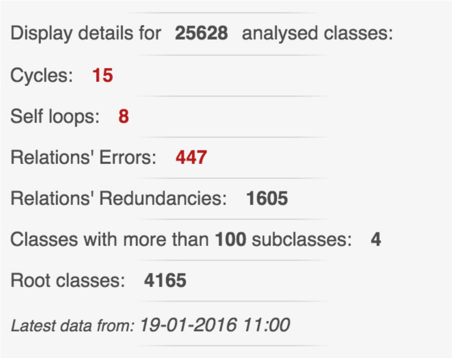
Relations' Errors: **450**

Relations' Redundancies: **1585**

Classes with more than **100** subclasses: **4**

Root classes: **3982**

Latest data from: 18-12-2015 17:12:11



Display details for **25628** analysed classes:

Cycles: **15**

Self loops: **8**

Relations' Errors: **447**

Relations' Redundancies: **1605**

Classes with more than **100** subclasses: **4**

Root classes: **4165**

Latest data from: 19-01-2016 11:00

Demo Presentation

Official launch – *October, 2015*

Users Feedback:

- *"Great tool! The error detection is precious!"*
- *"This is fantastic. :)"*
- *"Nice work! Thanks for sharing"*

`http://sergestratan.bitbucket.org`

All the information about the developed system, can be found on:
`https://bitbucket.org/sergestratan/sergestratan.bitbucket.org`

Conclusions & Future Work

Conclusions:

- Implemented a system for browsing the Wikidata taxonomy
- Provided some methods for the Wikidata ontology evaluation
- Applied different approaches to design and develop the system

Extensions:

- the Wikidata API integration
- extend ontology evaluation for additional quality criteria
- increase the amount of analyzed data

Thank You