Ontology issues in Wikidata

An overview

Why?

Using Wikidata as a source of knowledge requires effort

Often too much for small re-users

Ontology issues are a big problem for easy re-use of our data

- Unexpected query results and relations
- Inconsistency between similar concepts
- Even simple inferences are problematic, when connecting information

- Especially problematic for smaller and medium-size reusers

Approach

- Understand the current ontology issues (review of research and discussions)
- Figure out which ones are most important to address (discussions and survey?)
- Find ways to address the most important ones

Classification

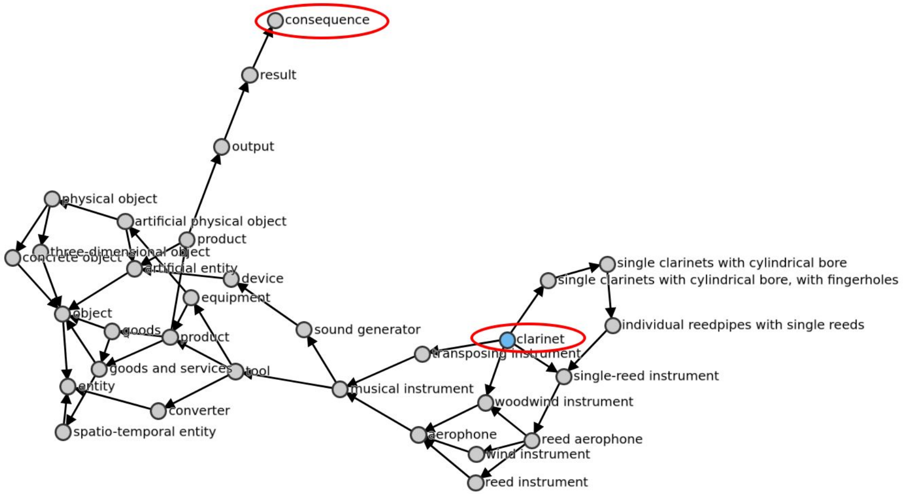
Overview of types of issues we found

- Semantic drift
- Structural bugs
  - cycles
  - mix-up of meta levels
  - redundant classification
  - redundant generalisation
  - exchanged sub-/superclasses

- Upper level ontology is messy
- Conceptual ambiguity
- Inconsistent modeling
- Overgeneralisation
- Conflicting real-world models

Semantic drift

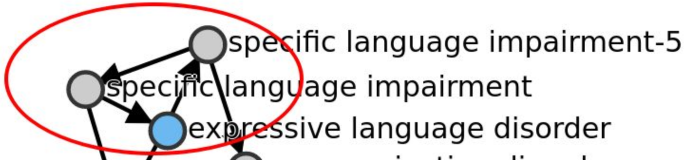
Super classes of “clarinet” (Q8343):

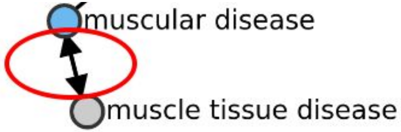


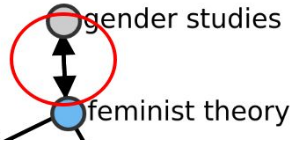
- "Subclass of" is assumed to be *transitive*: it holds between different levels of the class hierarchy
- Semantic drift shows when the inferences turn out to be wrong
- Individual subclass relations might be acceptable, but the combination is not.
- Caused by concepts having different aspects that are merged into one:
  - mason the person vs. mason the profession

Structural bugs

“subclass of” cycles


specific language impairment-5
specific language impairment
expressive language disorder

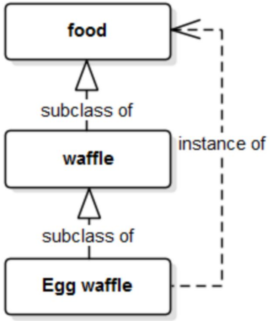


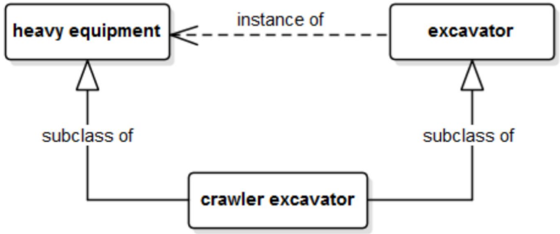


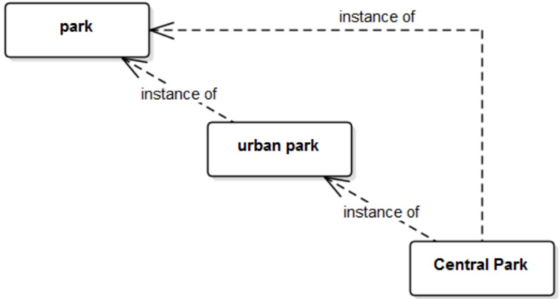
- Created if class A has a subclass B and B is a superclass of A
- Make it impossible to determine which Items are meant to be more specific or general than others
- Amounts to declaring that the classes A and B in a hierarchy are equivalent

Structural bugs

Mix-up of meta levels







- Occurs when, through inconsistent use of “instance of” vs. “subclass of”, the same Item is simultaneously a class and a metaclass, or similar.
- Brasileiro et al. (2016):
  - Z is both *instance of* and *subclass of* A
  - C has direct superclasses A and B such that B is *instance of* A
  - C is *instance of* both A and B, B is *instance of* A

Structural bugs

Redundant relations

Redundant Classification:

An Item is both an *instance of* a class and one of its super classes.

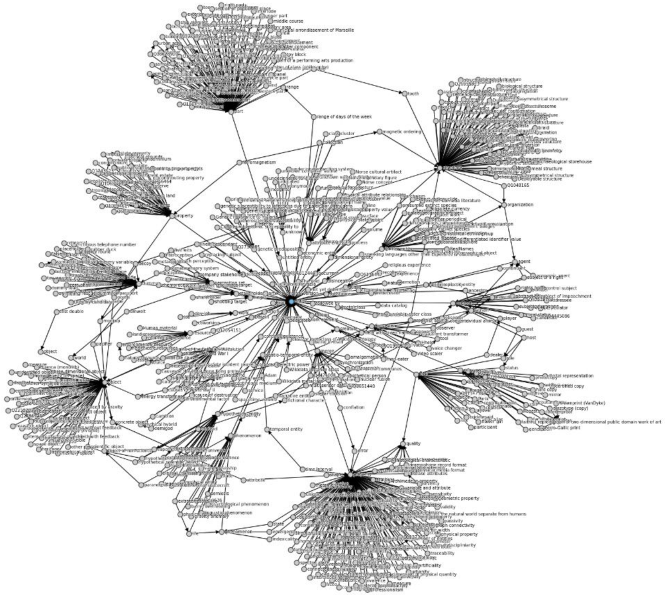
Redundant Generalisation:

An Item is both a *subclass of* a class and one of its super classes.

- If A is *instance of* B, which is *subclass of* C, then A *instance of* C is redundant
- If A is *subclass of* B, which is *subclass of* C, then A *subclass of* C is redundant

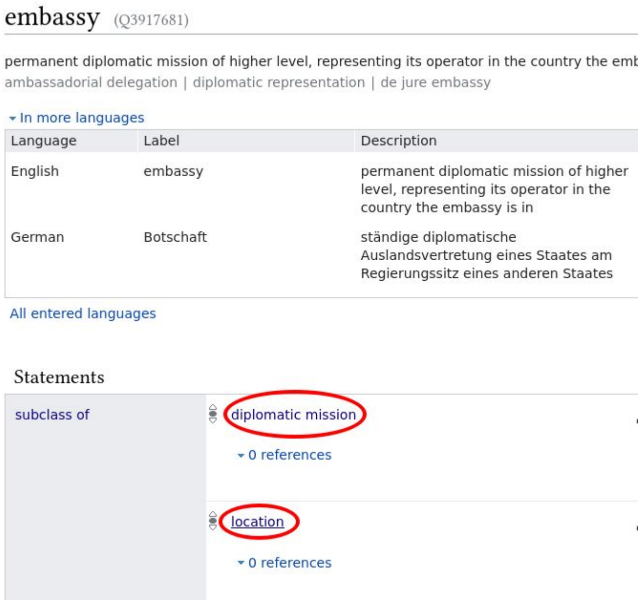
- Locality of editing: not seeing all the consequences of one's actions
- Potentially competing needs: sometimes the “shortcut statement” may be needed

Upper ontology is messy



- Upper ontology is hard™
- The top-class “entity” (Q35120) has 59 direct subclasses
- Messy connections in the upper ontology lead to:
  - issues with automated inferencing
  - nonsensical conclusions
- Do people care more about local ontologies?

Conceptual Ambiguity



embassy (Q3917681)

permanent diplomatic mission of higher level, representing its operator in the country the emb

ambassadorial delegation | diplomatic representation | de jure embassy

In more languages

Language	Label	Description
English	embassy	permanent diplomatic mission of higher level, representing its operator in the country the embassy is in
German	Botschaft	ständige diplomatische Auslandsvertretung eines Staates am Regierungssitz eines anderen Staates

All entered languages

Statements

subclass of

diplomatic mission

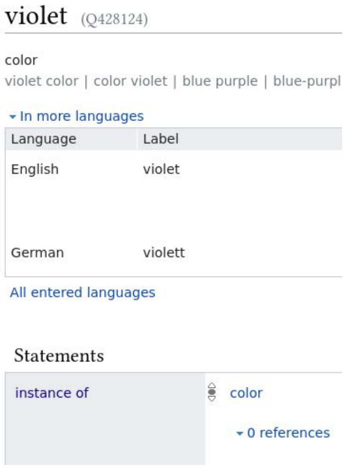
0 references

location

0 references

- Is caused by conceptual overloading of entities
- Makes it hard to understand what statements refer to
- Partly inherited from Wikipedia
- Partly created to integrate viewpoints
- Easier to keep overloading than to split (convenience)
- Alternative would be worse (significant increase in the number of Items)

Inconsistent Modeling



violet (Q428124)

color

violet color | color violet | blue purple | blue-purpl

In more languages

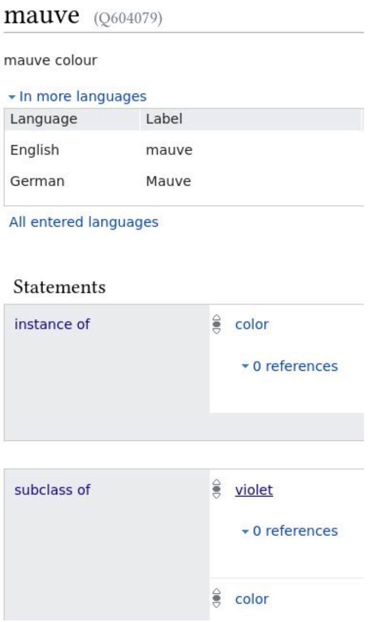
Language	Label
English	violet
German	violett

All entered languages

Statements

instance of — color

0 references



mauve (Q604079)

mauve colour

In more languages

Language	Label
English	mauve
German	Mauve

All entered languages

Statements

instance of — color

0 references

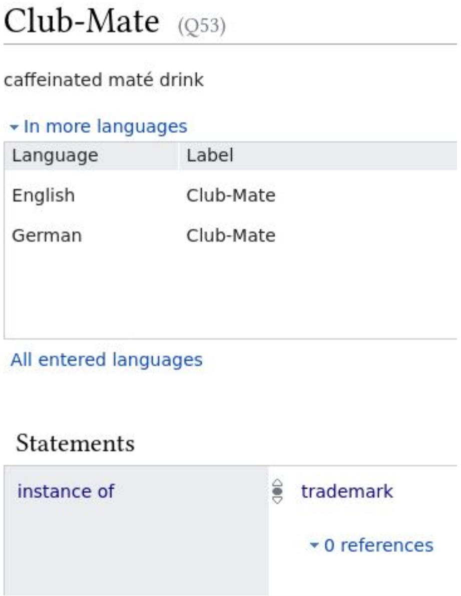
subclass of — violet

0 references

color

- Occurs when similar kinds of data is modelled in different ways
- Observable both across domains and within a single domain
- Example: mauve an *instance of* color and a *subclass of* one of its instances
    - What are colors?!
- Lack of common domain understanding?
- Several different ways to model the same data
- Very different design decisions taken for different domains

Over-Generalisation



Club-Mate (Q53)

caffeinated maté drink

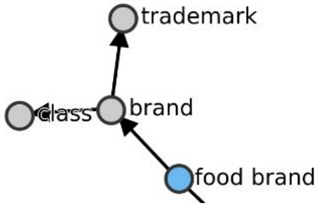
In more languages

Language	Label
English	Club-Mate
German	Club-Mate

All entered languages

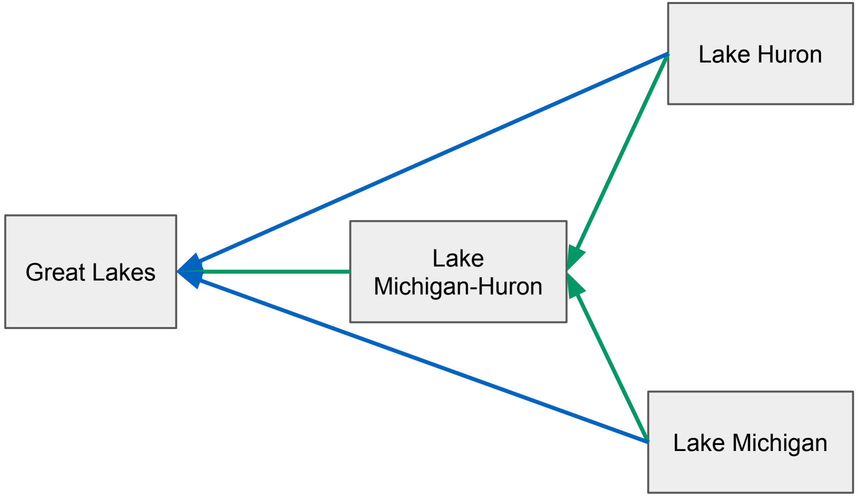
Statements

instance of	trademark 0 references



- Instances are too high in the class tree
- Classification is too general
- Example:
  - "Club Mate" (Q53) is a trademark, but it would be better classified as a "food brand", which is a "brand", which is "trademark", too.

Conflicting Real-World Models



- Real world is a mess
- Different groups have different views on the world
- May lead to overlapping and conflicting classifications
- Qualifiers to the rescue?

Questions

Questions

- Have you seen the issues presented?
- Can you think of any that are missing?
- Which ones are the worst?
- Why is everything so hard?
    - What is the source of those issues?
    - What's preventing them from being fixed already?
- What do you think would be helpful to have to address them?