Wikidata Ontology Issues.

Suggestions for prioritisation
based on the perceived frequency of occurrence
and the severity of impact on data re-use

Daria Ammalainen, UX Researcher

2023

Increase re-use for increased impact

In the Wikidata Development team, we believe that more people should be empowered to build applications using data from Wikidata.
To ensure that, among other initiatives, we want to work towards reducing ontology and data modelling issues.

In 2021 and 2022 we ran the Data Quality Days, which generated a lot of useful discussions on the processes around increasing/maintaining data quality and utility on Wikidata. As part of these discussions, we identified various types of ontology issues.

We wanted to explore the effect of the ontology issues on data reuse and identify the issues that have the most negative impact.

Research goals

Identify the most critical ontology issues

What types of issues complicate the development of apps and tools using data from Wikidata?

Explore the existing solutions and workarounds

How do data re-users currently deal with the ontology issues?

Update the classification of the ontology issues

Are there any other ontology issues that we don’t know about yet?

Method: Survey

To prioritise the most critical ontology issues, we conducted a survey.

In the main section of the survey, each of the previously identified **12 types of ontology issues were presented one by one** with a short description and an example of a problem.

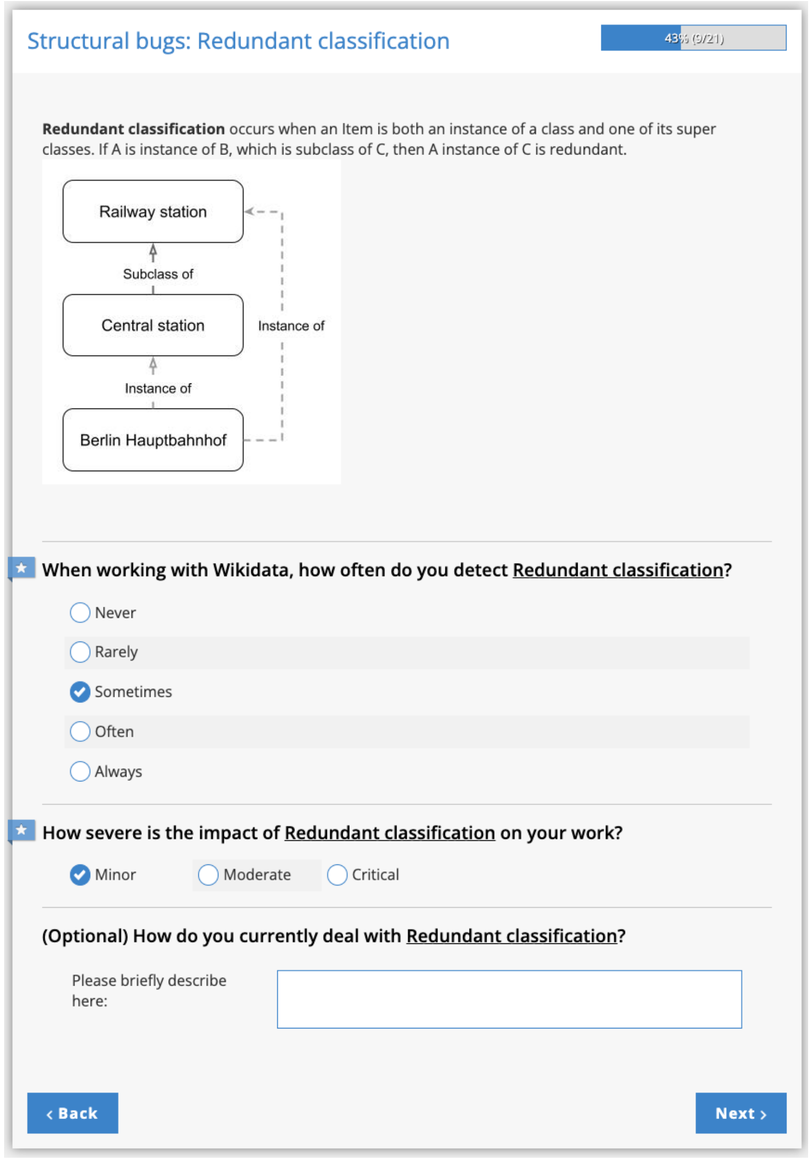
The participants were asked to estimate **how often they detect this type of issues** while working with Wikidata. Then they were asked to evaluate **the impact of these issues** on their work.

There was also an optional question on **current solutions and workarounds**.

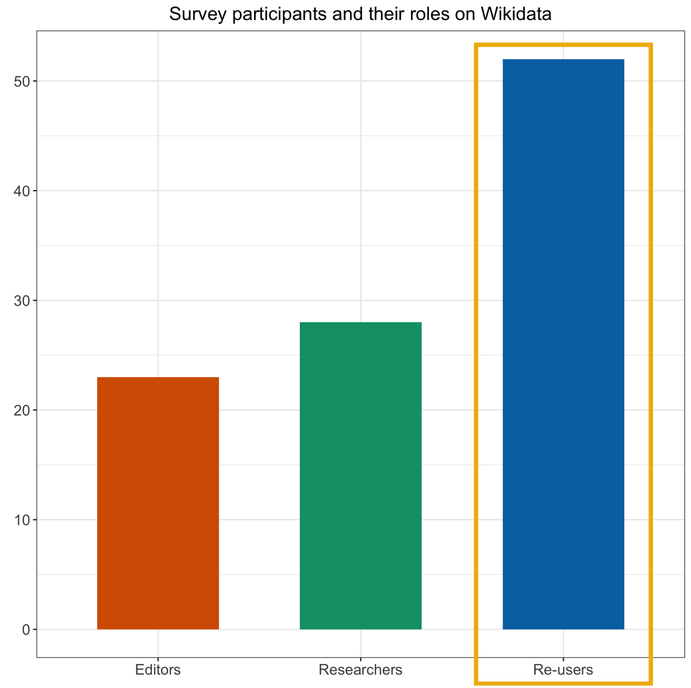
The participants could also share **the ontology issues missing from our classification**.

The last section of the survey covered the background information on the forms of activities on Wikidata.

The survey was announced on Project chat, Wikidata mailing list, Weekly summary, Wikidata:Ontology issues prioritization project page, Wikidata:WikiProject Ontology project page, and on Wikidata social media accounts.



Participants



Data re-users

As part of the initiative of empowering people to build applications and tools using data from Wikidata, in this research project we wanted to **focus on the experience of data re-users.**

For each of the issues, the frequency and severity ratings were calculated based on the median ratings of the participants identified as data re-users (N = 52).

Editors and Researchers

Other contributors identified in the survey include people editing Wikidata, importing data to Wikidata, and doing research using data from Wikidata. **The frequency and severity ratings from these types of contributors were not included in the analysis.**

However **the current solutions and workarounds, suggested missing ontology issues,** as well as the responses to the open-ended questions providing extra details and examples of the issues **were analysed combined with the responses from data re-users** (for those of you reading this — thank you so much for your detailed responses!).

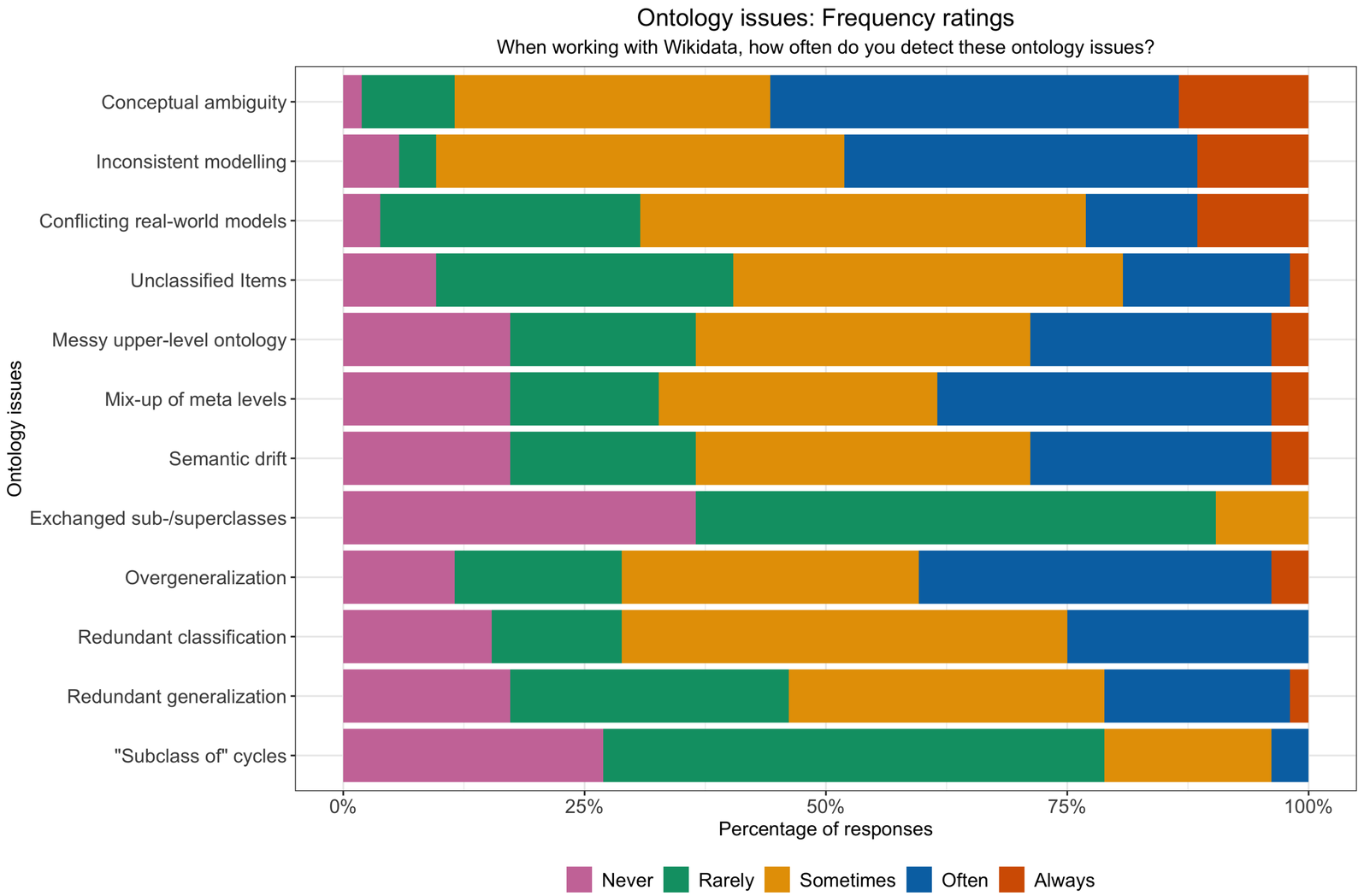
Most critical issues*

Conceptual ambiguity

Inconsistent modelling

*Issues with the highest median ratings of severity of impact on work and the highest median frequency ratings among data re-users.

Frequency ratings

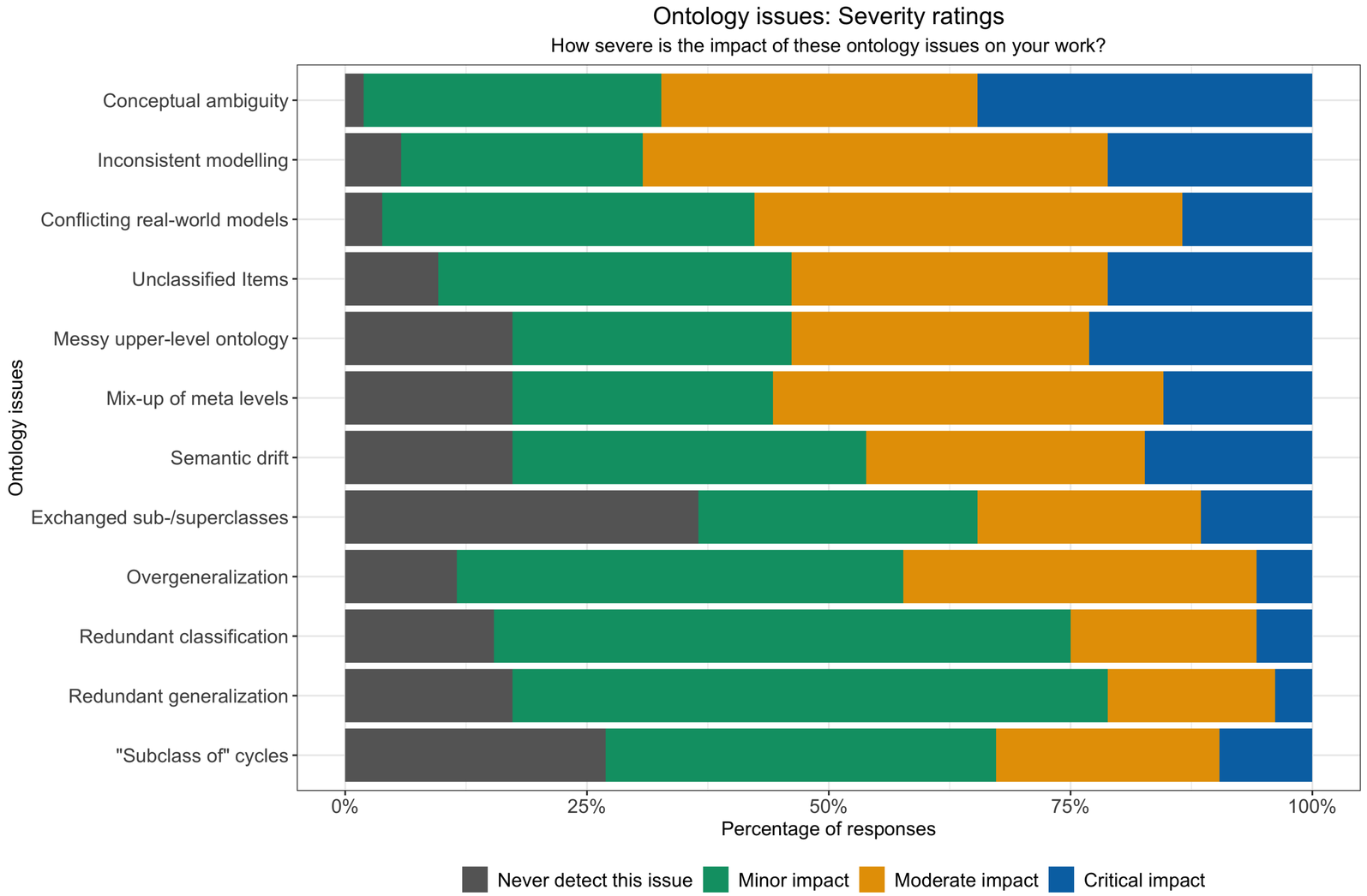


How often do you detect this issue?

The participants were asked to evaluate the frequency of each issue using a 5-point discrete frequency scale (from *Never* to *Always*).

The Frequency ratings are evaluated based on the responses attributed to data re-users. These are the responses from the participants, who indicated that they were building applications and tools using data from Wikidata (N=52).

Severity ratings



How severe is the impact of this issue on your work?

After evaluating the frequency, the participants assessed the impact of the issue on their work on a 3-step discrete severity scale (from *Minor* to *Critical*).

The participants who indicated that they never detected some of the issues, were not asked to evaluate the severity of impact of those issues on their work (the question was automatically skipped). These missing values are visualized in grey on the left side of the graph. This is done to make the visual comparison between the different issues more precise (and also for the graph to reflect the fact that some issues have no impact on some participants' work).

100% of responses on the graph N=52.

What exactly are those issues?

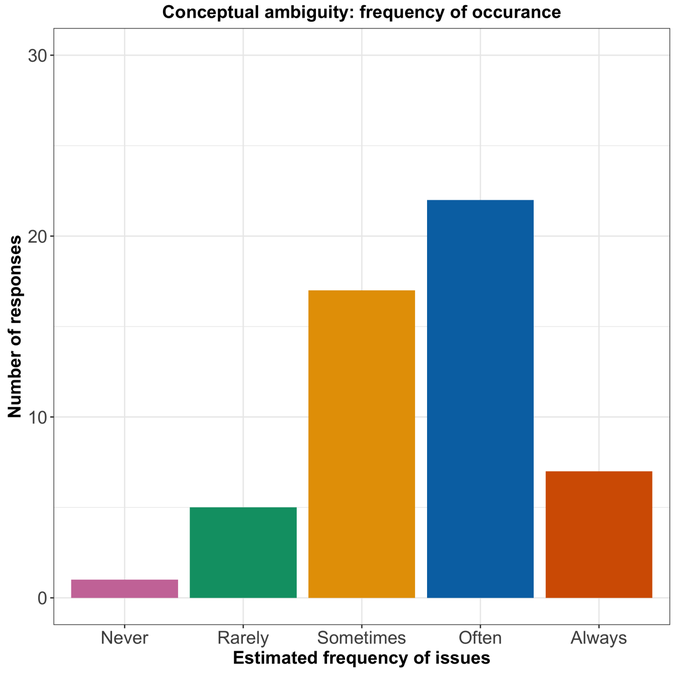
And how do people solve them?

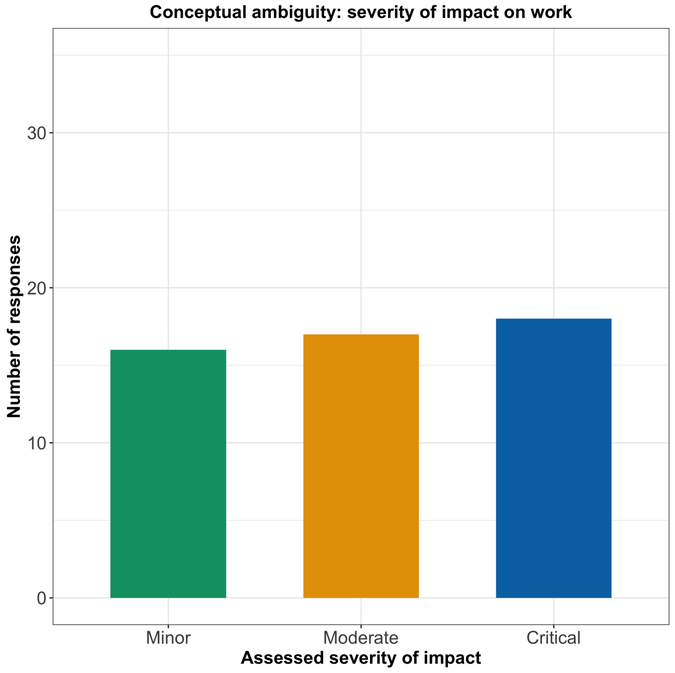
The next section explores the issues **in the order of descending severity of their impact and descending frequency of occurrence**, based on the median ratings from data re-users.

Each issue card includes the existing solutions and workarounds.

Conceptual ambiguity

Conceptual ambiguity happens when it is not clearly defined what an Item refers to. It is caused by conceptual overloading of entities. For example, the Item covers an embassy both as a location and a diplomatic mission. This makes it hard to understand what individual statements refer to.





The solutions and workarounds suggested by the participants include:

- Splitting the Items
- Editing the existing Item:
  - removing values to reduce the ambiguity
  - leaving the ambiguous values and adding tags to make the ambiguity more obvious
- Not using data from this part of Wikidata ontology
- Ignoring the problem:
  - the participants don’t have the domain expertise and choose not to intervene
  - the problem is too complex and there is no easy way to solve it individually
  - the issue is not frequent in the relevant domain

Received the highest median frequency and severity ratings among data re-users

Conceptual ambiguity: related issues

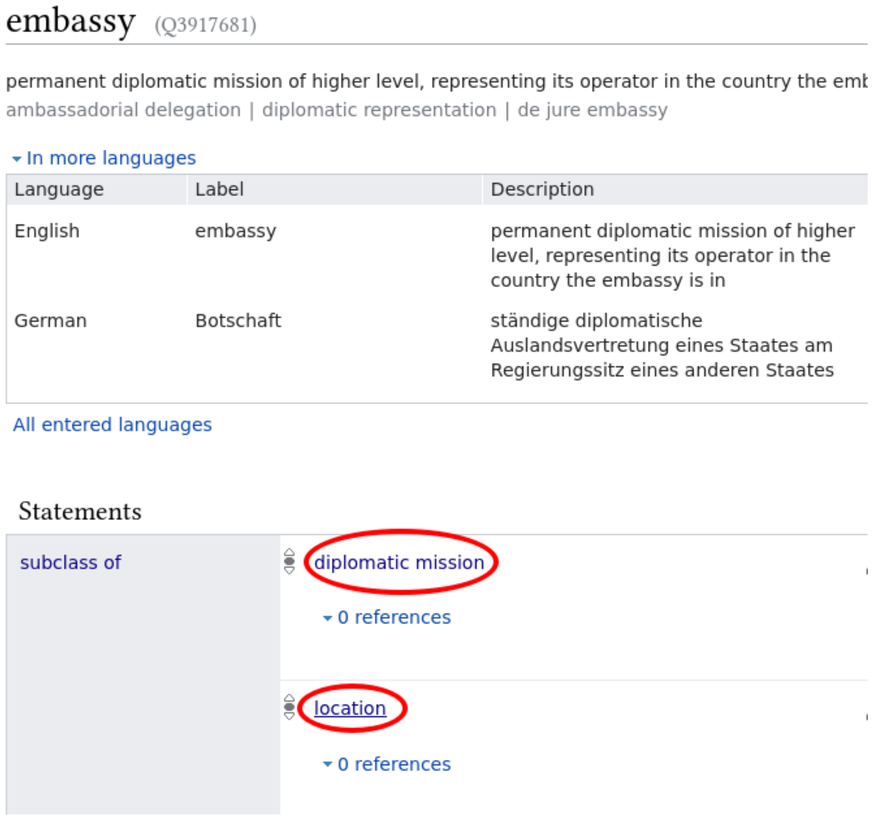
Conceptual ambiguity in properties and references.

These problems sometimes demotivate contributors to add and edit qualifiers.

Depending on the language, one Item might represent different concepts, so these differences might themselves cause the issues related to conceptual ambiguity or aggravate their impact on participants' work.

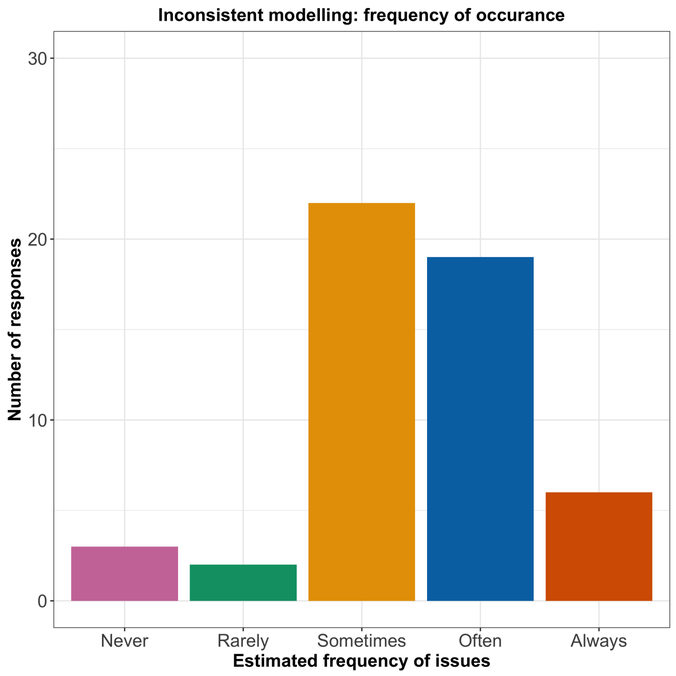
Cases when close/related (but not linked) **Wikipedia articles are matching different Wikidata Items** or when a Wikipedia article is matching several Wikidata Items.

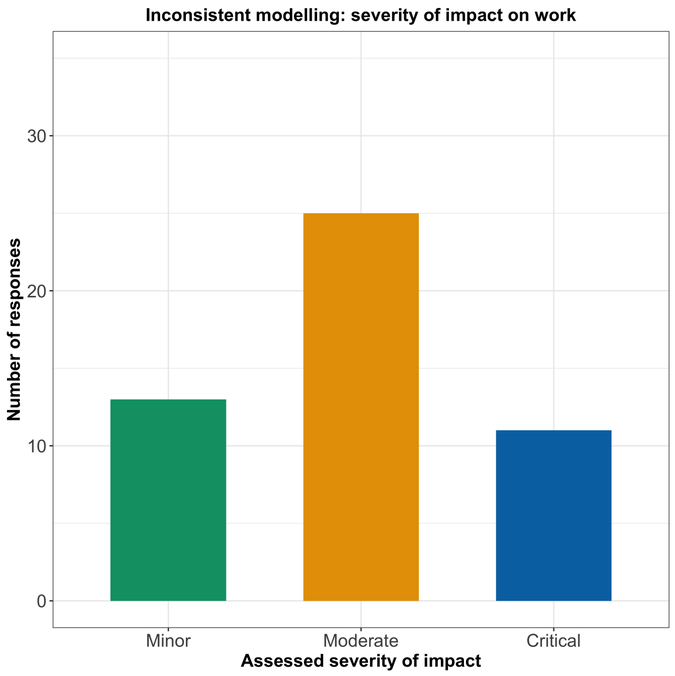
Some participants suggest that the underlying problem is the **difference in logic behind Wikidata and Wikipedia:** there will often be a single Wikipedia article for the entity, that is modelled by several Items on Wikidata.



Inconsistent modelling

Inconsistent modelling occurs when similar kinds of data are modelled in different ways.
It happens both across different domains as well as within a single domain.





Received the highest median frequency and severity ratings among data re-users

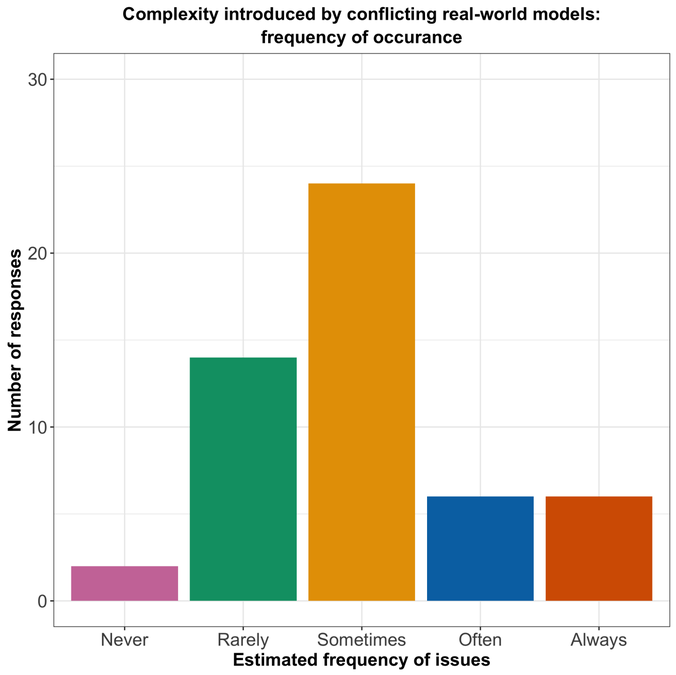
Current solutions and workarounds:

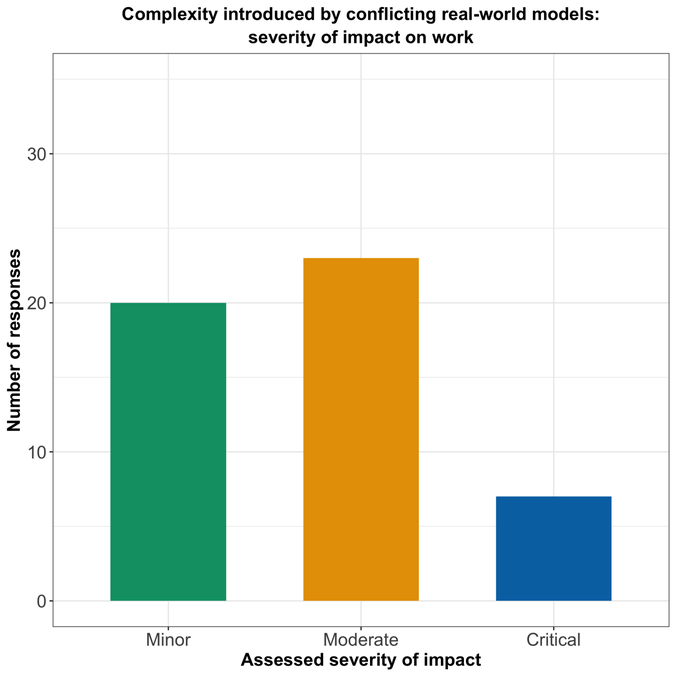
- Editing the Items:
  - finding and applying the model used in the relevant domain
  - adding references to existing claims
- Adjusting the queries:
  - to include / exclude data modelled differently
- Starting a discussion on talk pages (if the issue is widespread)
- Adding constraints
- Ignoring the problem

Participants also mentioned that it is difficult to solve the problem / to find a “right” way to model data, because the **help pages lack the important information or don’t exist at all.**

Complexity introduced by conflicting real-world models

The issues related to the **complexity introduced by conflicting real-world models** are caused by overlapping / alternative classifications of the same phenomenon. The different views on the world lead to different classification criteria and systems used in modelling.





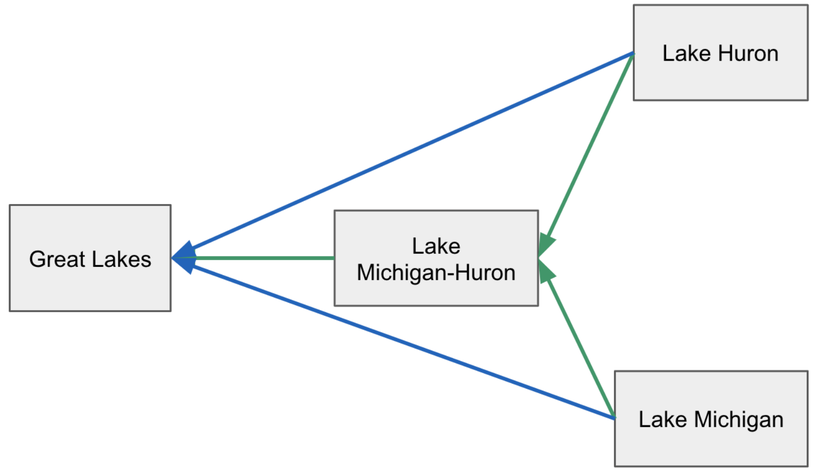
Existing solutions and workarounds:

- **Explicitly modelling the conflict / reflecting different viewpoints** !
- **Removing conflicting statements: applying one model** (e.g. that has a linked wiki project)
- Starting a discussion on relevant talk pages
- Excluding / removing the data from the export
- Including the data in the export and documenting or visualising the way the data is modelled
- Ignoring the issue:
  - the problem is too complex
  - the problem originates from Wikipedia versions, and it is not clear how to solve it in Wikidata
  - lack of domain expertise
  - the issue is not frequent in the relevant domain

Conflicting real-world models

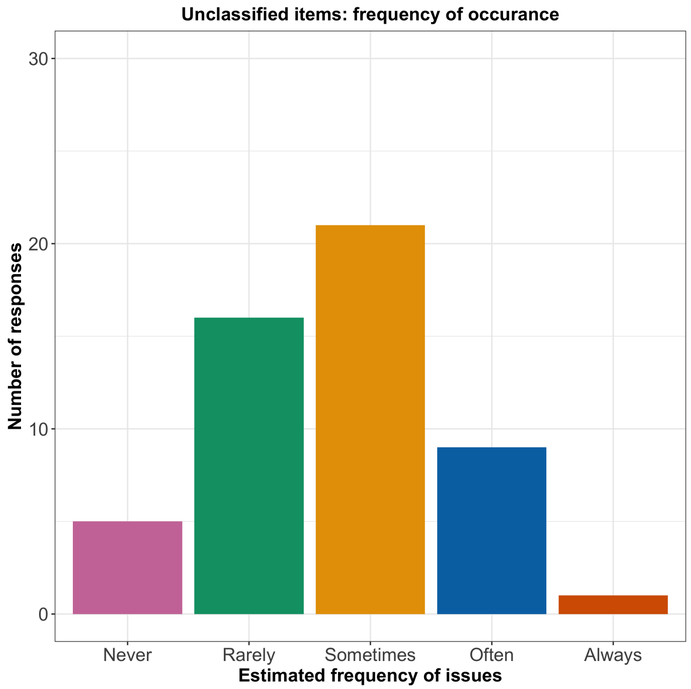
Some participants suggest that the **general ontological plurality is a feature of Wikidata, but there should be some supporting mechanisms** to better deal with different views on the world.

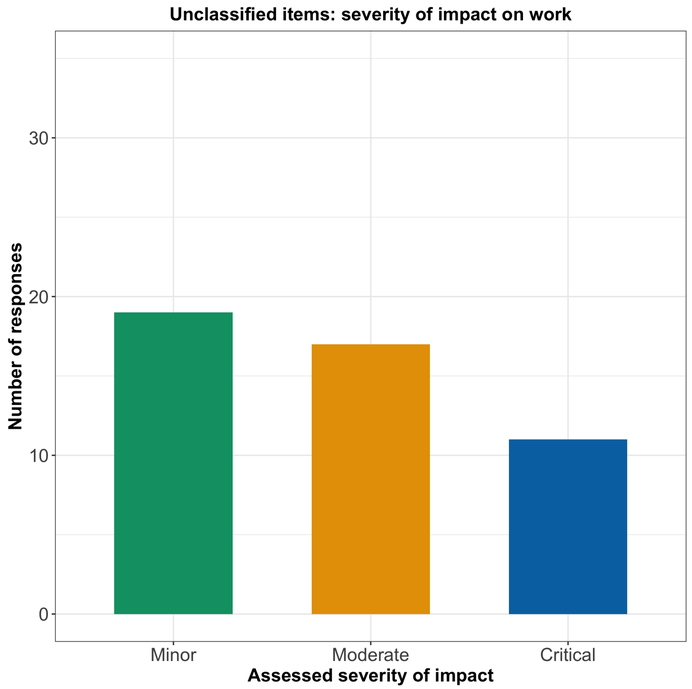
It was also highlighted, that **it is difficult to deal with this issue** both in terms of data reuse and mass import **without the subject matter expertise and / or guidance.**



Unclassified items

A number of Items have no classifying statements (e.g. “instance of”, “subclass of”, “part of”) and are therefore not connected to the existing ontology. This means they will not show up in certain query results among other things.



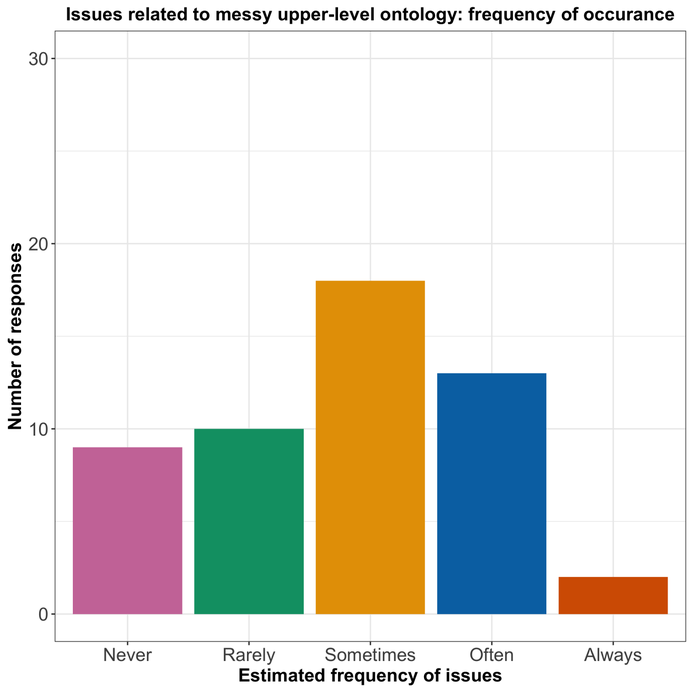


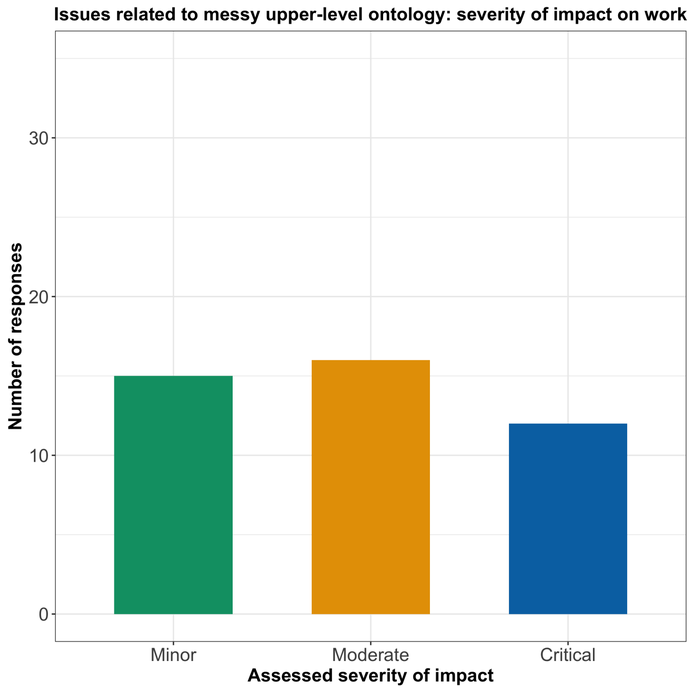
The existing solutions and workarounds include:

- Adding statements to classify the Items
  - including using Psychic to predict P31 and P279, and PetScan for mass edits
- Not using this part of Wikidata ontology (if the problem is frequent in the domain of interest)
- Ignoring the problem

Messy upper-level ontology*

The highest level of Wikidata's ontology contains many connections. These connections are sometimes arguably wrong, conflicting or too detailed. **Messy connections in the upper ontology** may lead to nonsensical conclusions and issues with automated inferencing.





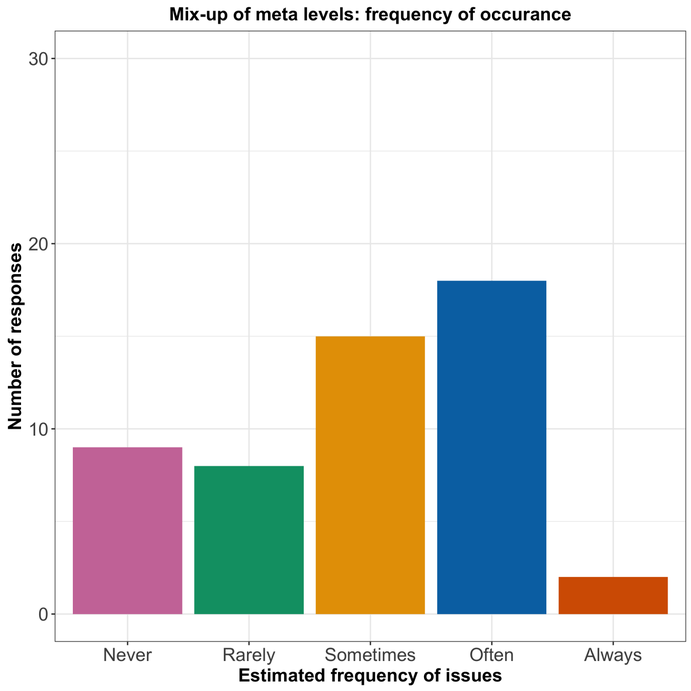
The solutions and workarounds include:

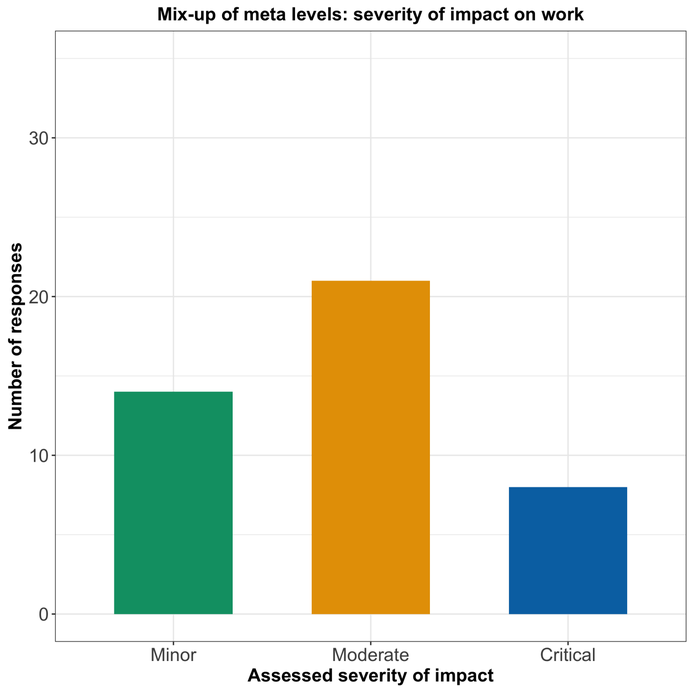
- Using only small subsets of data
- Starting a discussion on relevant talk pages
- Editing the items (applying the model used in the relevant domain)
- Ignoring the problem:
  - too broad / to complex to solve individually
  - it is not affecting their work

**note: at least some of the participants might have evaluated the frequency and severity of impact of the messy connections in the upper level of the part of ontology in their domain area of interest, rather than the global upper-level ontology (suggested by the analysis of the responses to the open-ended question about current solutions). That might have affected the median frequency and severity rating of this issue.*

Mix-up of meta levels

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.



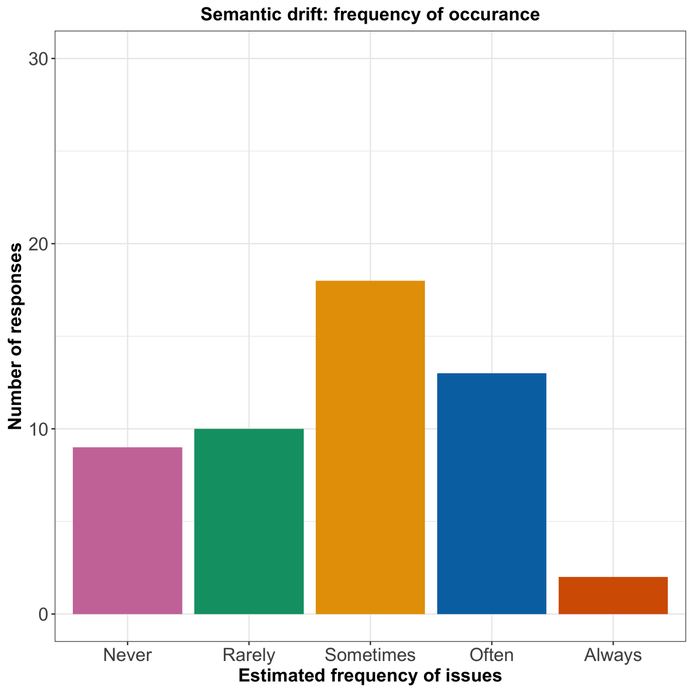


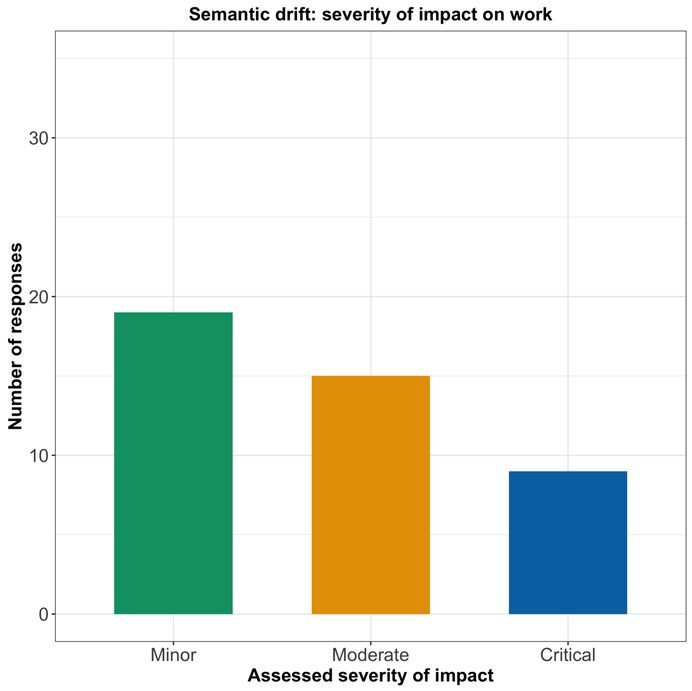
The suggested solutions and workarounds include:

- Editing the Items:
  - following the model used in the relevant domain
  - supporting claims with external ontology
  - removing “instance of” and leaving only “subclass of”
- Adjusting the queries (to include / exclude the data)
- Data cleanup (after the export)
- Ignoring the issue / waiting for the solution

Semantic drift

"Subclass of" is assumed to be transitive, meaning it holds true between different levels of the class hierarchy. **Semantic drift** shows up when inferences are wrong because they assume this transitivity. It happens when the concepts with different aspects are combined in one Item (e.g. mason the person vs. mason the profession), which can lead to wrong inferences.



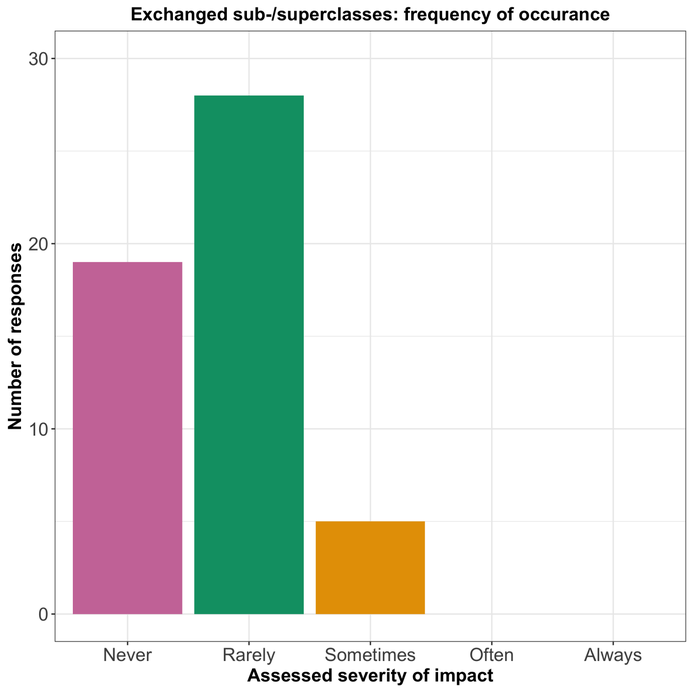


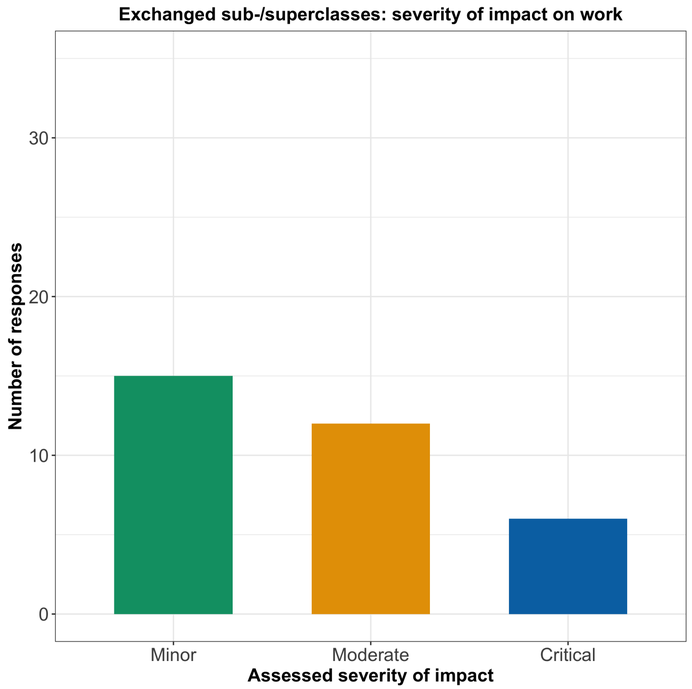
Current solutions and workarounds include:

- Splitting the Items
- Starting a discussion on relevant talk pages / reporting the issue to the community
- Manually finding the source of the problem and editing (e.g. removing the subclasses leading to the problem)
- Data cleanup (after the export)
- Not using the affected part of Wikidata ontology
- Ignoring the issue

Exchanged sub-/superclasses

There are cases where the ***subclass* and *superclass* are switched**, leading to a wrong relation in the class hierarchy.



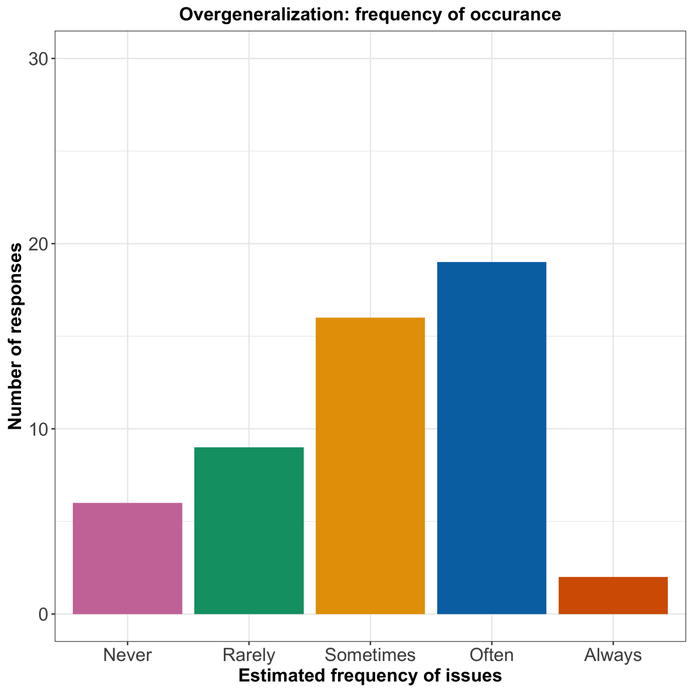


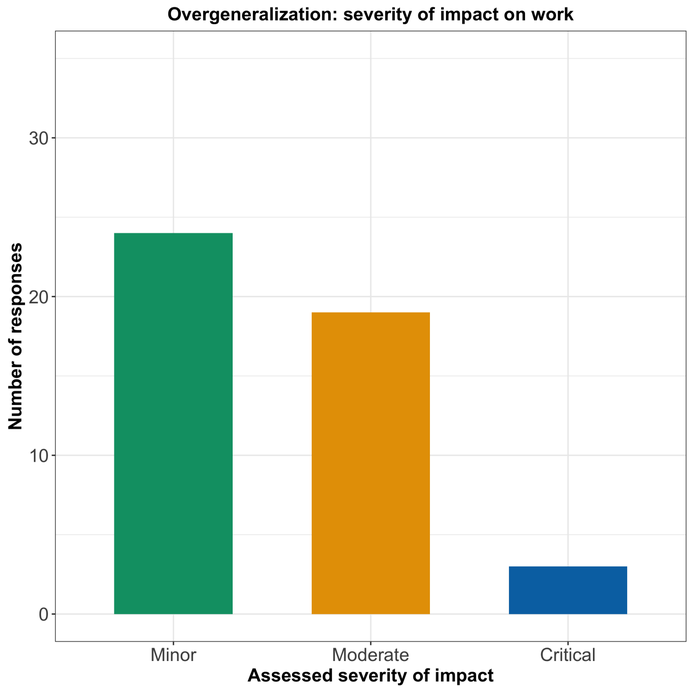
The current solutions and workarounds include:

- Editing: switching sub- and superclasses
- Adding references to existing statements
- Ignoring the issue

Overgeneralization

Overgeneralization happens when instances are too high in the class tree.
This means the classification of some entities is too general.



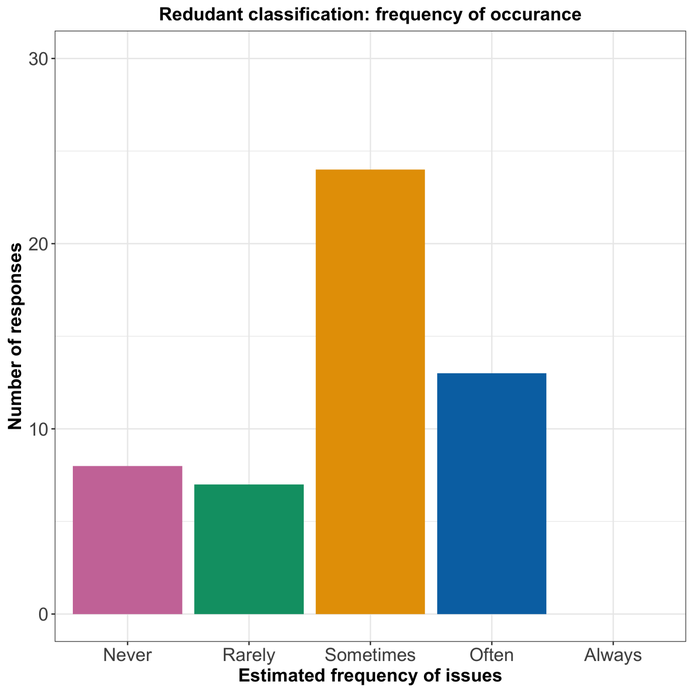


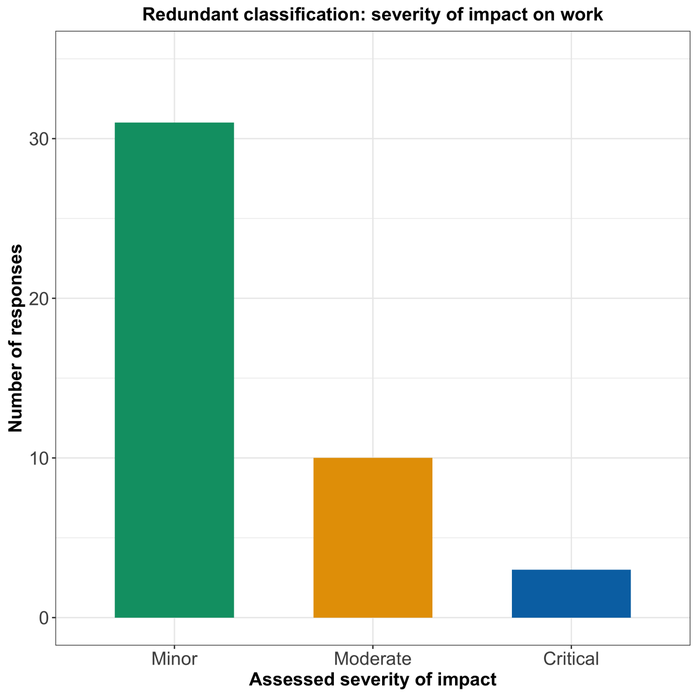
The solutions and workarounds include:

- Editing the Items: changing the class to a more specific one
- Adjusting queries to include all relevant data
- Ignoring the problem
  - including treating the issue as a natural maturity flow / case of missing information

Redundant classification

Redundant classification occurs when an Item is both an instance 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.



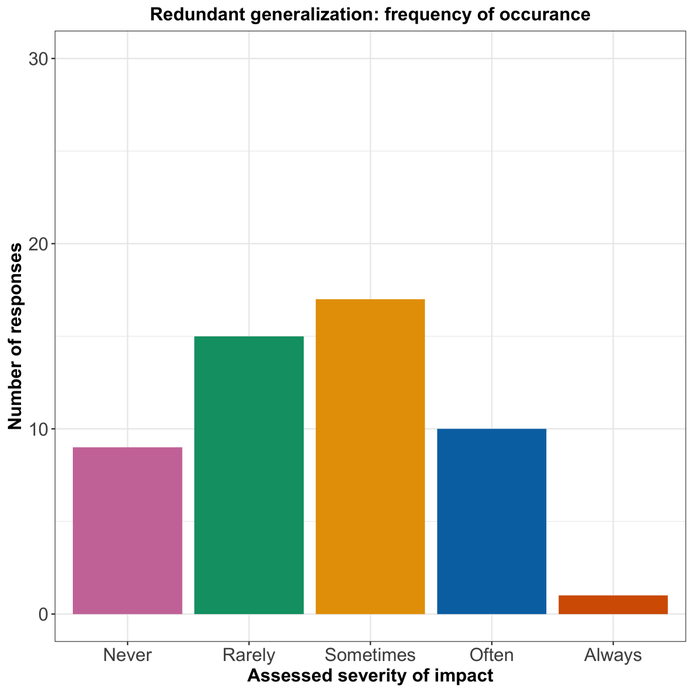


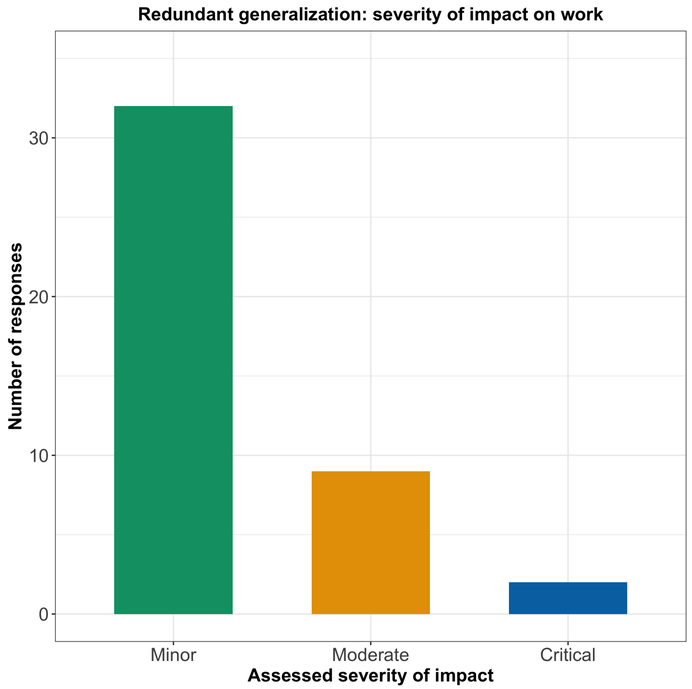
The current solutions and workarounds:

- Editing the Items: removing more generic statements
- Adding references to existing statements
- Starting a discussion on relevant talk pages
- Adjusting queries to exclude the data
- Data cleanup after the export
- Ignoring the issue:
  - it is too broad / complex to solve individually
  - it is not affecting their work
  - they do not perceive it as a problem

Redundant generalization

Redundant generalization occurs when an Item is both a subclass of a class and one of its super classes. If A is *subclass of* B, which is *subclass of* C, then A *subclass of* C is redundant.



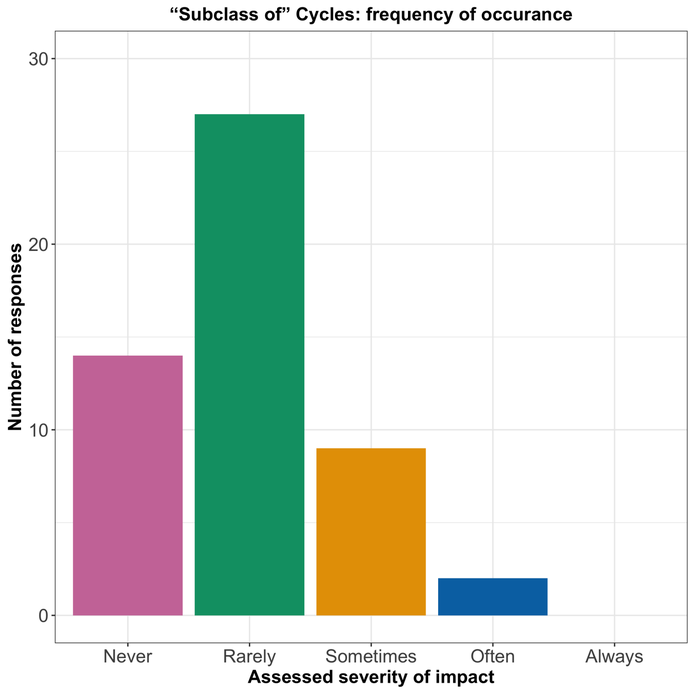


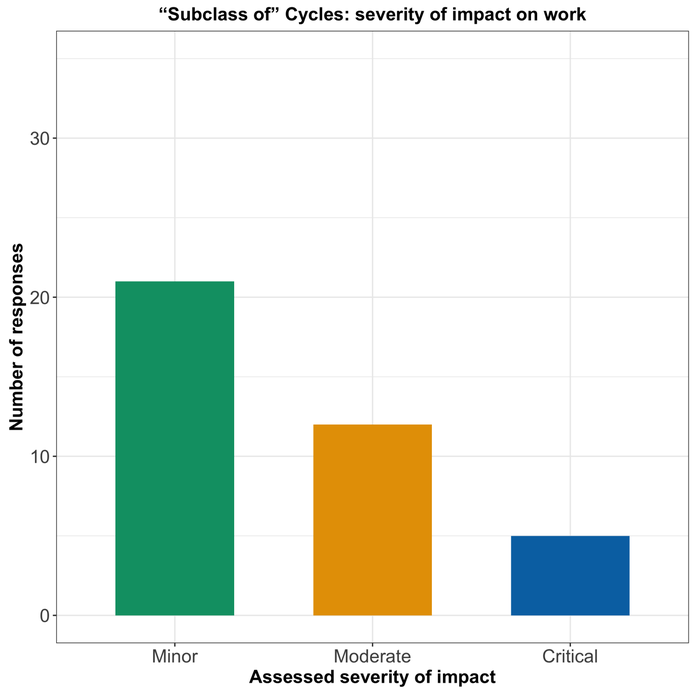
The current solutions and workarounds include:

- Editing: removing the more generic statement
- Adding references to existing statements
- Data cleanup after the export
- Adjusting queries to include / exclude the data
- Ignoring the issue: it has minor impact on their work / is not perceived as a problem

“Subclass of” cycles

“Subclass of” cycles are created if class A has a subclass B and B is a superclass of A. These cycles make it impossible to determine which Items are meant to be more specific or general than others.





The current solutions and workarounds include:

- Editing:
  - removing one of the subclasses
  - following a model used in the relevant domain
- Adding references to existing statements
- Not using this part of Wikidata ontology
- Ignoring the issue

Some participants also reported **“part of” cycles**, which have a similar underlying issue.

Suggested additional ontology issues

The participants suggested other ontology issues that were missing from the classification.

Duplicate Items of the same entities

- lead to messy ontology
- might be the result of mass imports

Classes with too many direct subclasses

- might be the result of bot activity

Other problems

📌 There are not a lot of resources on best practices for data modelling on Wikidata

The most precise properties are difficult to identify without the domain expertise or the examples to copy from.

Some participants suggest that a best practices page or supporting tool would help them classify the Items and solve the ontology issues when they find them while working with Wikidata.

📌 Wikidata's ontology is not stable and solutions to ontology issues have to be constantly updated

This sometimes leads to data re-users switching to a different knowledge-bases or working with only a subset of Wikidata.

Other topics brought up

- The whole Wikidata ontology cannot be viewed (only pages covering domain-specific branches of Wikidata ontology)
- Inconsistent constraint messages
- Item Completeness

Thank you! What's next?

Discussing the survey results

Please share your thoughts and comments
at Wikidata_talk:Ontology_issues_prioritization

Identifying the approaches to addressing the ontology issues