Evidence of Large-Scale Conceptual Disarray in Multi-Level Taxonomies in Wikidata

Atílio A. Dadalto [a], João Paulo A. Almeida [a,*], Claudenir M. Fonseca [b], Giancarlo Guizzardi [b]
[a] *Ontology and Conceptual Modeling Research Group (NEMO), Federal University of Espírito Santo, Brazil*
E-mail: atilio.dadalto@aluno.ufes.br; jpalmeida@ieee.org
[b] *Services & Cybersecurity Group, University of Twente, The Netherlands*
E-mail: c.moraisfonseca@utwente.nl; g.guizzardi@utwente.nl

Abstract. The distinction between types and individuals is key to most conceptual modeling techniques and knowledge representation languages. Despite that, there are a number of situations in which modelers navigate this distinction inadequately, leading to problematic models. We show evidence of a large number of representation mistakes associated with the failure to employ this distinction in the Wikidata knowledge graph, which can be identified with the incorrect use of *instantiation*, which is a relation between an individual and a type, and *specialization* (or *subtyping*), which is a relation between two types. The prevalence of the problems in Wikidata's taxonomies suggests that methodological and computational tools are required to mitigate the issues identified, which occur in many settings when individuals, types, and their metatypes are included in the domain of interest. We conduct a conceptual analysis of entities involved in recurrent erroneous cases identified in this empirical data, and present a tool that supports users in avoiding some of these mistakes.

Keywords: Wikidata, Multi-Level Taxonomies, Evaluation

1. Introduction

Types are predicative entities, whose instances share some general characteristics, i.e., they are said to be repeatable invariances across multiple individuals. Individuals (or tokens), in their turn, are not general sorts of things, they are not repeatable; instead, they are particular entities, like Paul McCartney and John Lennon (instances of "person") or Jupiter and Mars (instances of "planet"). While we seem to be able to grasp this distinction intuitively, the boundaries between types and individuals are not always sharply drawn in everyday discourse. Consider, for instance, the paradigmatic case of "word" [1]. How many words are there in the sentence "the book is on the table"? The answer is *six* if we count the two occurrences of "the" as distinct words (or word tokens), or *five* if we count the word *types* used in the sentence. When we say "they drive the same car", do we mean the same *type of car* (qualitative identity) of the same *individual car* (nummerical identity)?

Given its occurrence in natural language, it is not surprising that this kind of ambiguity can arise also in knowledge representation and conceptual modeling. For instance, if we are capturing invariants about the domain of cars, what kinds of properties will characterize an entity named "car"? An *individual car* has a chassis number and a production date, while a *type of car* (or car model) can be characterized by the tag sales price, set of available colors, etc.

*Corresponding Author: Av. Fernando Ferrari, 514, Vitória, ES, Brazil

Distinguishing between these two interpretations is key to grasp what an instance of "car" stands for, and what kinds of relations it can establish with other entities in a model. An instance of *type of car* can specialize another type of car, in the way that "Porsche Speedster 23F" specializes "Four-Wheeled Car". An instance of *individual car* can instantiate "Porsche Speedster 23F", in the way that James Dean's Porsche did.

This paper examines the use of this distinction in practice, by employing Wikidata as a source of empirical data. Wikidata is structured as a graph with millions of nodes called *items*. A Wikidata item may represent a type (class) (e.g., the item for `planet (Q634)`) or an individual (e.g., the item for `Earth (Q2)`). The edges of this graph represent relations between items including specialization and instantiation. We here uncover a large number of items whose relations to other items indicate that their interpretation as a type or as an individual may be ambiguous. The prevalence of the problems in Wikidata's taxonomies suggests that guidelines are required to mitigate the large-scale conceptual disarray identified. Special attention is required to the so-called multi-level taxonomies, when meta-types (such as `astronomical object type (Q17444909)`) are represented in tandem with (first-order) types and individuals (and meta-meta-types, etc.). In these multi-level taxonomies, the confusion occurs not only between individuals and (first-order) types, but also between types of adjacent successive orders. Some of these problems were originally identified by some of us in [2] and characterized in terms of a number of anti-patterns, i.e., recurrent error-prone model structures; we now revisit two of these anti-patterns here in further detail, following several years of changes in Wikidata.

After eliciting this empirical data, we then conduct an analysis of a number of entities in Wikidata that are frequently involved in these anti-patterns. We identify some of the possible reasons behind these violations and, by using logical, ontological and semantic considerations, we propose some possible interpretation solutions for eliminating them. Finally, we demonstrate how we can leverage on these anti-patterns to build automated procedures that can proactively detect these violations before they are introduced to Wikidata. This paper is an extended version of a short paper published in [3]. The short paper only covered one of the anti-patterns we cover here. Further, we also provide here evidence that attempts to introduce a multi-level mechanism present in OpenCyc [4] did not address the issues in multi-level taxonomies in Wikidata. Overall, there is more in-depth discussion on the multi-level problems uncovered.

This paper is further organized as follows: Section 2 discusses how Wikidata supports (multi-level) taxonomies. It shows some problems that occur when instantiation and specialization are combined in the platform. Section 3 identifies these problems at scale, updating some of the statistics collected in the 2016 for Wikidata [2]. Section 4 examines these results in an attempt to identify a conceptual basis for explaining the identified problems, as well as proposing possible interpretation solutions for rectifying them. Section 5 presents a Web application that illustrates how the anti-patterns exemplified on these problems can be proactively detected before they are introduced in Wikidata. Finally, Section 6 presents final considerations, including related work.

2. Taxonomies in Wikidata

Knowledge in Wikidata consists of *statements* that capture relations between *items*, which are "*are used to represent all the things in human knowledge*" [5]. A statement has the form of a "<subject> <property> <object>" triple. Examples of widely-used properties include `instance of (P31)` and `subclass of (P279)`. The property `instance of (P31)` represents a relation between an instance and a class (i.e., type), where the latter is predicated of the former. For example, `Earth (Q2)` is an `instance of terrestrial planet (Q128207)`, therefore exhibiting the properties of that class, in this case, being a planet of mostly rocky and metallic composition. The property `subclass of (P279)`, on the other hand, holds between two classes where the subclass has as instances a subset of the instances of the superclass. For example, `terrestrial planet (Q128207)` is a subclass of `planet (Q634)` meaning that every instance of the former is also an instance of the latter.

Wikidata also allows the declaration of classes of classes (or meta-classes). For example, `terrestrial planet` is `instance of` the class `astronomical object type (Q17444909)`, whose instances are specializations of `astronomical object (Q6999)` (see Figure 1). The work of [6] clarifies this scheme of classes stratified in meta-levels (i.e., class, meta-class, meta-meta-class), using the concept of order, where individuals (entities that cannot have instances, like `Earth`) instantiate first-order classes, who in turn instantiate second-order classes, and so on

into orders above (e.g., third-order, fourth-order). Figure 1 presents this reiterated application of `instance of` relations forming a multi-level taxonomic structure using the items mentioned above. Here boxes represent items, while dashed arrows and solid arrows represent `subclass of (P279)` and `instance of (P31)`, respectively.

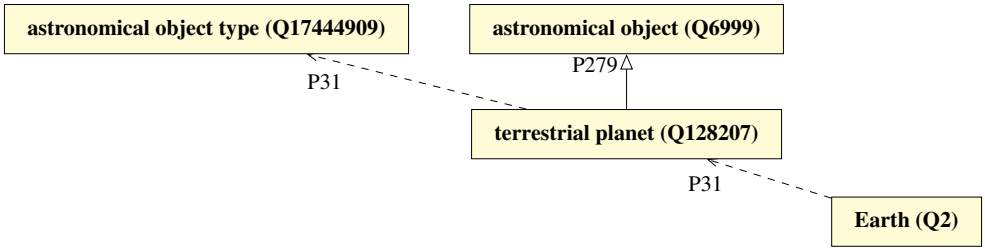

Fig. 1. Terrestrial planet: instance of astronomical object type, subclass of astronomical object

Other types of astronomical objects are also present in the platform, such as `star (Q523)`, which, again, is an instance of `astronomical object type` and subclass of `astronomical object`. In this domain, there is a clear stratification into individuals (such as `Earth`, `Alpha Centauri (Q12176)`), first-order types (such as `planet`, `star`), and a second-order type (`astronomical object type`). Note that we retain the capitalization of labels and plural forms from Wikidata.

This same clear stratification is not present in other taxonomic structures of Wikidata, however. Consider, for instance, the following fragment concerning the French language, depicted in Figure 2. `French (Q150)` is both *instance of* and *subclass of* `language (Q34770)`. This opens up multiple interpretations: is French meant to be referring to a *type* of language or a specific, *particular* language? Of course, it is known that the French language is a particular language that has a certain number of speakers at a given point of time; however, variants of that language have spawned over the years, which can be considered instances of a class of French languages. The same ambiguity applies to these variants, such as `American French (Q3083193)`, which denotes the "varieties of the French language that are spoken in North America". The two facets (language as a *class* and language as a *particular*) are confounded in Wikidata.

Another example of difficulty in classification is found in the representation of colors, as shown in Figure 3. Again, through the usage of both instantiation and specialization, it is unclear whether `Turkey red (Q3443194)` is a *particular color* or a *kind of color*. Furthermore, we see that `Turkey red` is subclass of both `red (Q3142)` and `color (Q1075)`. This implies that all instances of `Turkey red` are instances of `color` and `red`. Meanwhile, however, instances of `red` cannot be instances of color since `red` itself is already an instance of `color`. In fact, this

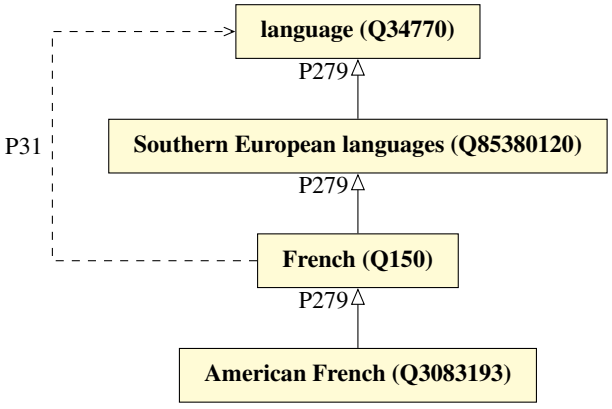

Fig. 2. French as instance and subclass of language.

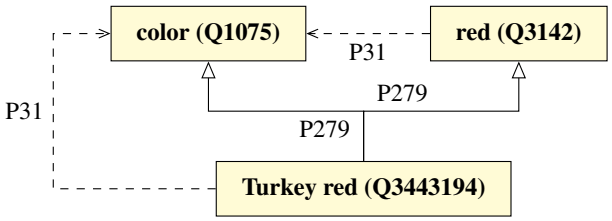

Fig. 3. Turkey red as a specialization of color and of its instance red.

part of the model seems to mix up the notions of *color region* (e.g., red) in the color spindle and *color points* (i.e., atomic regions designating a super-determinate color shade) [7].

3. Assessment of Taxonomic Structures in Wikidata

The problems identified in the previous section are instances of recurring patterns involving instantiation and specialization that were identified originally in [2]. More precisely, they are instances of *anti-patterns*, since they represent error-prone recurrent structures. The fragment exemplified by the French language is called here anti-pattern 1 (AP1 for short), and occurs whenever an item is instance of and subclass of another item (direct or indirectly) at the same time. AP1 prevents stratification into orders since, at the one hand, instantiation forces related items to be at different adjacent orders, and, at the other hand, a specialization of a class at a certain order must be in that same order (for formalization of the underlying theory and proofs, see [6, 8]). The fragment involving colors is called here anti-pattern 2 (AP2 for short), and occurs whenever stratification into orders is prevented by an item being a subclass of two items, one of which is an instance of the other. In this section, we discuss how we detect those patterns at scale in Wikidata and summarize the data we collected in the platform.

3.1. Data Collection

In order to deal with the size of Wikidata, we used a filtered dump of the Wikidata database[1] as of 14 September 2020. Because our interest is only on taxonomic structures, we have selected only statements with entities declared as subclasses (i.e., that have the `P279` property asserted). The dump was created using wdumper[2] and processed using Stardog 7.4 and Jena 4.0.0. It has 2,452,006 entities, 26,264,034 statements and 38,224,283 triples, roughly 2.5% of the almost 100,000,000 entities present in the complete Wikidata database as of April 2021.

3.2. Anti-Pattern Occurrences

To assess the occurrence of the anti-patterns, we have executed SPARQL queries in the filtered dump. Listing 1 shows the SPARQL query used to find AP1 occurrences considering transitiveness for *subclass of* statements.

Listing 1: SPARQL query for AP1.

```
SELECT DISTINCT ?subject ?class WHERE {
  ?subject wdt:P31 ?class .
  ?subject wdt:P279+ ?class .
}
```

[1] https://zenodo.org/record/4046102
[2] Further dump details and mirrors at https://wdumps.toolforge.org/dump/749.

We have found 2,035,434 `?subject ?class` pairs involved in AP1, covering domains such as biology, gastronomy, awards, professions, sports, among others.

Regarding the second anti-pattern, we have obtained 3,006,945 results. Listing 2 shows the SPARQL query for AP2 with transitivity for subclassing. Due to computational reasons, we limited transitivity in AP2 to a maximum of 8 levels. Queries for AP2 were executed using Apache Jena 4.0.0 as there were performance issues with Stardog.

Listing 2: SPARQL query for AP2.

```
SELECT DISTINCT ?subject ?class1 ?class2 WHERE {
  ?subject wdt:P279+ ?class1 .
  ?subject wdt:P279+ ?class2 .
  ?class2 wdt:P31 ?class1 .
}
```

Transitivity of subclassing is important as it reveals a large number of anti-pattern occurrences, which could indicate that it is harder to identify the specialization paths to indirect superclasses. The AP1 query without subclassing transitivity (P279) yields 1,279,629 results, while a query considering transitivity (P279+) returns 2,035,434 results. This finding is substantially more pronounced with AP2: when there is no transitivity in the AP2 query, only 646 results are returned, in contrast to 3,006,945 results when transitivity up to 8 levels is introduced.

3.3. Entities Most Frequently Involved in Anti-Patterns AP1 and AP2

We have produced a ranking of the entities most frequently involved in the anti-patterns so that they could be further analyzed. The 20 top-ranked entities involved in AP1 are listed in Table 1 along with the number of times it participates in the anti-pattern. A comprehensive ranking with 200 entities and all scripts used in this paper are available at https://purl.org/nemo/wapa.

Table 1
Ranking of occurrences of entities involved in AP1.

Place	Wikidata QID	English label	AP1 occurrences
1	Q7187	gene	971,982
2	Q8054	protein	757,360
3	Q4164871	position	103,545
4	Q277338	pseudogene	49,404
5	Q427087	non-coding RNA	49,132
6	Q2996394	biological process	30,315
7	Q12136	disease	12293
8	Q14860489	molecular function	11,204
9	Q34770	language	6,795
10	Q5058355	cellular component	4,287
11	Q294414	public office	2,544
12	Q898273	protein domain	2,493
13	Q282	wine	2,143
14	Q929833	rare disease	1,994
15	Q618779	award	1,469
16	Q55788864	developmental defect during embryogenesis	1,403
17	Q201448	transfer RNA	1,153
18	Q11173	chemical compound	875
19	Q60754876	grade of an order	772
20	Q55789477	head and neck disease	735

There is a clear overlap of subdomains in the ranking, especially but not limited to those entities related to biology and biochemistry, e.g., gene as a "*basic physical and functional unit of heredity*" and pseudogene (Q277338) as a "*functionless relative of a gene*". For example, gene is a well-known multi-faceted concept frequently referring to a particular gene type repeatable in each chromosome of every cell (gene instances, i.e., particular biochemical structures composed of particular nucleotides) but also to the representation of a gene type (a data object) that results from genome sequencing operations. For both gene and pseudogene, multiple anti-pattern occurrences involving them are introduced from batch adding or merging statements from external knowledge databases such as UniProt and NCBI Gene, without proper consideration of whether the imported entities are types or individuals. Hundreds of thousands of genes are directly related to gene (Q7187) in immediate instantiation and specialization relations! This pattern repeats for instances of protein, protein domain, disease, rare disease, development defect during embryogenesis, head and neck disease, non-coding RNA, transfer RNA. Users and softbots alike leverage databases such as GeneDB (genes), UniProt (proteins) Disease Ontology (diseases), InterPro, PubMed, NCBI Gene (RNAs), Gene Ontology (biological processes, cellular components), thus, introducing these violations. Other domains highly present in AP1, include social roles and titles (e.g., position (Q4164871), public office (Q294414), award (Q618779), and grade of an order (Q60754876)), language classification (e.g., language (Q34770)), and products of controlled origin denomination (e.g., wine (Q282)).

We inspected some of these top entities in the ranking to identify in which exact revision in the history of the Wikidata updates a violation was introduced. For example, take language (Q34770). Originally, the item Guarani (Q35876) was simply represented as being an *instance of* language. However, revision 174811757 introduced the statement that Guarani (Q35876) is a *subclass of* indigenous language of the Americas (Q51739)—which is an indirect *subclass of* language (Q34770). Together these statements configure a case of anti-pattern AP1. An anti-pattern checker could play a role in this context by detecting revisions that introduce inconsistencies prior to the inclusion of new statements.

Table 2 shows the top 20 entities for AP2. A large number of entities that appear in the ranking for AP1 also appear here: biological process (Q2996394), position (Q4164871), disease (Q12136), etc.

Table 2
Ranking of occurrences of entities involved in AP2.

Place	Wikidata QID	English label	AP2 occurrences
1	Q2996394	biological process	627,925
2	Q4164871	position	400,141
3	Q12737077	occupation	287,711
4	Q12136	disease	222,823
5	Q28640	profession	192,386
6	Q14860489	molecular function	54,513
7	Q294414	public office	41,198
8	Q11862829	academic discipline	39,505
9	Q16889133	class	28,801
10	Q1207505	quality	21,544
11	Q5058355	cellular component	18,162
12	Q2424752	product	18,051
13	Q4936952	anatomical structure	15,129
14	Q11028	information	8,539
15	Q55788864	developmental defect during embryogenesis	7,014
16	Q33104279	philosophical concept	6,483
17	Q781413	mental process	6,314
18	Q130901	binary relation	6,121
19	Q18123741	infectious disease	5,991
20	Q1914636	activity	5,884

3.4. Anti-Patterns Statistics Considering The OpenCyc Basic Scheme

Although Wikidata is in principle 'level-blind' [9], i.e., in its basic item scheme it does not include leveling mechanisms, the platform includes a set of classes representing types of different orders, namely `first-order class (Q104086571)`, `second-order class (Q24017414)`, `third-order class (Q24017465)`, `fourth-order class (Q24027474)`, `fifth-order class (Q24027515)`, and `fixed order metaclass of higher order (Q24027526)`. These classes are declared as equivalent to their counterparts in the OpenCyc ontology [4]. Hence, it is possible to analyze the occurrences of anti-patterns under this basic 'OpenCyc scheme'. It can be verified for the analyzed dump that the scheme is not widely used in Wikidata; e.g., by querying for instances of classes that specialize `fixed-order metaclass (Q23959932)`, as shown in Listing 3, only 178 instances of fixed-order classes are found.

Listing 3: SPARQL query for fixed-order classes.

```
SELECT ?class WHERE {
    ?class wdt:P31 ?fixedOrder .
    ?fixedOrder wdt:P279 wd:Q23959932 .
}
```

Listing 4 shows the SPARQL query for obtaining AP1 occurrences present concomitantly with the OpenCyc scheme.[3]

Listing 4: SPARQL query for AP1 with OpenCyc.

```
SELECT ?metaclass ?subject ?fixedOrder WHERE {
    ?subject wdt:P31 ?class .
    ?subject wdt:P279+ ?class .
    ?class wdt:P31 wd:Q24017414 % Q24017465, etc
}
```

By querying for AP1 while considering the OpenCyc layer, it is found that, despite efforts to lay stratification into rigid orders, there are still a large number of anti-pattern occurrences to be found even when topmost entities involved are placed as fixed-order classes, under the OpenCyc scheme. There are 770,638 occurrences of AP1 with classes at the top explicitly marked as instances of a `fixed-order metaclass (Q23959932)`, which translates to 37.9% of all occurrences of AP1. Take, for example, `computer science (Q21198)`: not only it is simultaneously subclass and instance of `academic discipline (Q11862829)`, but `academic discipline (Q11862829)` is also an instance of `second-order class (Q24017414)`.

The results above show that fixing entities to strict orders does not remove all forms of anti-patterns from Wikidata. This might be due to the fact that many real-world concepts have been modeled into Wikidata using heterogeneous modeling notions, if at all, which makes it difficult to fit these entities into rigid orders. However, considering that a significant amount of human effort goes into editing content on Wikidata, merely posing entities as fixed-order classes isn't enough to stave off modeling problems, since (i) proper classifications do not prevent users from creating contradictory statements and (ii) Wikidata encompasses users from every background and they might not be acquainted with Cyc (or any modeling schemes), rendering these models devoid of meaning for most of them. For these reasons, it is important to find ways of dealing with anti-patterns in practice, rather than relying exclusively upon formal, abstract analysis; we need to consider ways of tackling these issues without resorting to a priori knowledge about how entities are modeled and should relate to one another. Moreover, the violation of the OpenCyc model illustrates that more than just its usage is necessary to avoid anti-patterns. To this end, in Section 5 of this

[3] Due to performance reasons, data regarding AP2 could not be collected.

paper, we present a tool for users of Wikidata to identify occurrences of AP1, capable of analyzing the state of Wikidata and also the implications of introducing new, hypothetical statements.

4. Analysis and Discussion

The top-ranking entity involved in the anti-patterns we investigated is `gene`, which is described in Wikidata as a "basic physical and functional unit of heredity" with instances such as `TP53 (Q14818098)`, a "protein-coding gene in the species Homo sapiens". Inspecting their use in Wikidata, instances of `gene` like `TP53` are most likely not "a particular gene from one cell from one person" but instead a *type* of which "many of us have tokens of — in fact many tokens of in each cell of our bodies" [10]. There is evidence for this in the properties ascribed to `TP53`, such as "found in taxon Homo Sapiens" and "encodes Tumor protein p53". This is consistent with an interpretation of `gene` as a second-order class, and its instances (e.g., `TP53`) as first-order classes. However, `TP53`, besides being declared as an instance of `gene`, is declared a subclass of `protein-coding gene (Q20747295)`, which is itself a subclass of `gene`. Therefore, `TP53` (and most of the other instances of `gene`) is also a subclass of `gene`. How should instances of `TP53` be interpreted then, as they are also instances of `gene` like `TP53` itself? We hypothesize that the subclassing statement is incorrect. `TP53` is — not a subclass of, but—an instance of the `protein-coding gene (Q20747295)` subclass of `gene`. This issue may have never been flagged in Wikidata as instances of instances of gene are never instantiated explicitly in the platform (as it is not tracking "a particular gene from one cell from one person", but types of these). In fact, most gene talk is quantifying over types as discussed by Wetzel [10]. The same observation can be made for the other entities in the ranking related to biology and biochemistry such as: protein, pseudogene, non-coding RNA, cellular component, rare disease, development defect during embryogenesis, transfer RNA, chemical compound, and head and neck disease. These are all second-order types whose instances are first-order types whose instances are not recorded in the platform. Hence, there is a mismatch between ontological considerations (`TP53` is instantiated in a particular cell in a Petri dish, and, hence, `TP53` is a class) and knowledge representation considerations (items instantiating `TP53` are never recorded in Wikidata).

Further in the ranking, there are related entities such as `position (Q4164871)` (in the sense of "*social role [...] within an [...] organization*") and its subclass `public office (Q294414)`. An instance of `position` is `mayor (Q30185)`, "head of municipal government such as a town or city", instantiated by `Frank Hilker (Q104772317)`. Clearly, he is an individual! Hence, `mayor` is a first-order class, suggesting `position` is a second-order class. However, `mayor` is declared as a subclass of `public office` which is a subclass of `position`. As a consequence, we come to the absurd inference that `Frank Hilker` is an instance of `position` (and consequently an instance of its superclasses, like `artificial entity (Q16686448)`)![4]. We hypothesize the declaration of `mayor` as a subclass of `position` is incorrect. The former being a first-order class and the latter a second-order class. As discussed in [6], order-crossing specialization is logically incorrect. Differently from the case of gene, the platform includes instances of instances of `position` (such as `Frank Hilker`); similarly, though, `gene` and `position` are second-order classes (meta-classes). It is important to note here that Wikidata has a specialized property to declare occupation of a position by a person (`position held (P39)`) and this is used instead of instantiation for most declarations of occupation. In any case, one needs to settle whether `mayor` and other entities like this are instances or specializations of `position` irrespective of the use of `position held`.

The case of `biological process (Q2996394)` also reveals confusion in the identification of the order for that entity. It is a subclass of `process (Q3249551)`, which in turn is a subclass of `occurrence (Q1190554)`, which is then described as "occurrence of a fact or object in space-time". An occurrence may be qualified by `point in time (Q186408)`, which is indicative that its instances are individual occurrences. Hence, `biological process` should be considered a first-order class. However, `biological process` includes among its instances entities such as `birth (Q14819852)` and `death (Q4)`, entities bearing their own instances. The latter has as instance the `death of James Dean (Q15213260)`. Hence, death is a class of biological processes, and we must conclude—contra our earlier conclusion—that `biological process` should be considered a second-order class, as `death` is

[4] In [2], some of us have shown that Tim Berners-Lee was inferred to be an instance of `profession` due to the same anti-pattern; this is no longer the case in the current state of the platform.

not an individual, but a type. Here we note that although biological process type (Q47989961) exists as an item, it is not used to classify birth, death, etc.

The case of language, which we have raised earlier, involves the representation of extremely rich phenomena with much variation and diversity (a spectrum of macrolanguages, language families, dialects). In this case, the criteria for individuation for a language is difficult to establish, and, as discussed earlier, items such as French can be regarded as a particular language or as a class of similar languages (given that each of its variations may be considered itself a language). We should note that language is an instance of languoid class (Q28923954) (described as "e.g. dialect, language, macrolanguage, language subfamily, family, or superfamily; each instance of these is a subclass of languoid"). And, languoid class is explicitly marked as second-order class in Wikidata (it is an instance of Wikidata metaclass (Q19361238) which is an instance of third-order class (Q24017465)). This makes language a first-order class, and its instances individuals. As individuals, instances of language must not be involved in *subclass of* statements. To separate the two facets of a language, we need two items: one representing the language (say French of France (Q3083196)) as an instance of language (or dialect), and another as a subclass of language (or dialect) (referring to the class of French variants, whose instances include Quebec French (Q979914), Swiss French (Q1480152), and French of France).

The case of wine (Q282) may be indicative of a problem in establishing a criteria of individuation for its instances. Take, for example, Italian wine (Q1125341), a subclass of wine, and Rosso di Montalcino (Q25993), an instance of wine and a subclass of Italian wine. In this excerpt, either, (i) the instantiation of Rosso di Montalcino is incorrect and its subclassing is correct, and therefore, all three of them should be considered types at the same order, or (ii) the instantiation of Rosso di Montalcino is correct and its subclassing is incorrect, in which case wine should in fact be considered a type at a higher order. Option (i) is consistent with wine being a subclass of alcoholic beverage (Q154) which is an instance of type of food or dish (Q19861951), a second-order class. It is further consistent with the issues discussed for gene and the other biochemical entities: the platform is not in the business of recording information about particular portions of wine, which may make it hard for its users to "anchor the definition" in a level of individuals. Wine may be particularly challenging because it is a noun that takes on countable and uncountable usage (as a mass expression). "[W]hen we switch from speaking of 'wine' to 'a wine' or 'seven wines', we usually switch from speaking about wine, or portions of it, to speaking about kinds of wine" ([11] *apud* [12]).

Finally, there are the cases of award (Q618779) and grade of an order (Q60754876). A variety of awards are given periodically, such as the Academy Awards, the Pulitzer Prize, and the Turing Award. Many of these awards are claimed to be, simultaneously, instances and subclasses of award. For example, this is the case of the item for the well-known Emmy Award (Q123737). Like biological process, award is also an indirect subclass of occurrence (Q1190554). Hence, if considered a subclass of award, Emmy Award should be a first-order class, and its instances particular occurrences (like the granting of the "Emmy Award for Outstanding Lead Actress – Miniseries or a Movie" in 1978 to Meryl Streep). The same happens with grade of an order. For example, its instance Commander of the Order of Orange-Nassau (Q1861904) is also a subclass of commander (Q524980) which is a subclass of grade of an order. Here again, there seems to be a confusion between the use of instantiation and subclassing.

Note that the rankings for both anti-patterns we have presented in this paper have been filtered to remove entities that are marked as instances of variable-order class (Q23958852), since these are explicitly flagged as not being stratified into a particular order. Variable-order [4] (or orderless [8]) classes have instances at different orders. Thus, being an orderless class can justify its occurrence in the anti-patterns without incurring in an error of classification. This is the reason why these classes have been excluded from our analysis.

5. Automated Support

By leveraging on the type of analysis conducted in the previous section and the anti-patterns that can be identified with it, one can implement automated procedures for proactively identifying occurrences of these anti-patterns before they are introduced in Wikidata. In this section, we illustrate that by implementing such a procedure for the

case of AP1 as a Web application termed the Wikidata Anti-Pattern Analyzer (or WAPA for short)[5]. WAPA allows the user to input any entity from Wikidata to check for existing occurrences of AP1, or input full hypothetical statement to verify whether it would introduce new violations. Since it retrieves data directly from Wikidata's SPARQL endpoint, the results reflect the current state of Wikidata (in the screenshots below, they reflect the state of Wikidata in April 2021).

To illustrate its usage, let's take `French (Q150)` as an example, as shown in Figure 2. If we input `French (Q150)` and check for existing anti-patterns, the tool will correctly return the fact that, currently, `French (Q150)` is simultaneously instance and subclass of `language (Q34770)`. It will also look for violations within its instances and subclasses. `French (Q150)` has dozens of subclasses but no instances, hence WAPA reports that there are no entities that are simultaneously instances and subclasses of `French (Q150)`.

An example of violation in subclasses and superclasses of a single entity is seen with `mining industry (Q1945600)`, in Figure 4, where it is, simultaneously, instance and subclass of `industry (Q268592)` (April 2021). Also, `mining of metal ores (Q16638398)` is, simultaneously, instance and subclass of `mining industry (Q1945600)`. Conflicting statements like these make it difficult to pinpoint an entity's position in a taxonomy, and this tool can assist users to detect violations.

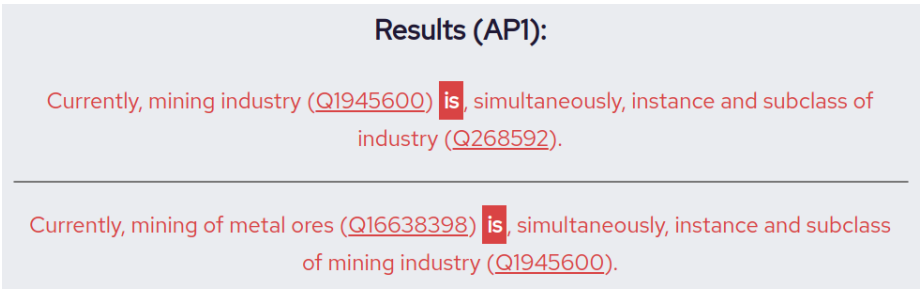

Fig. 4. WAPA results when checking for violations regarding `mining industry (Q1945600)`.

WAPA can also analyze the validity of hypothetical statements. For example, if a user wants to state that `Pulitzer Prize (Q46525)` is `subclass of science award (Q11448906)`, this tool can check whether the inclusion of this statement would introduce violations to Wikidata. Indeed, in this case, `Pulitzer Prize (Q46525)` would be, simultaneously, instance and subclass of `science award (Q11448906)`. Since WAPA always checks for existing violations before testing the hypothetical scenario, it would also return that `Pulitzer Prize (Q46525)` is, simultaneously, instance and subclass of `journalism prize (Q1709894)` in addition to the results for the hypothetical statement.

[5] Accessible in https://atilioa.github.io/WikidataAntiPatternAnalyzer/.

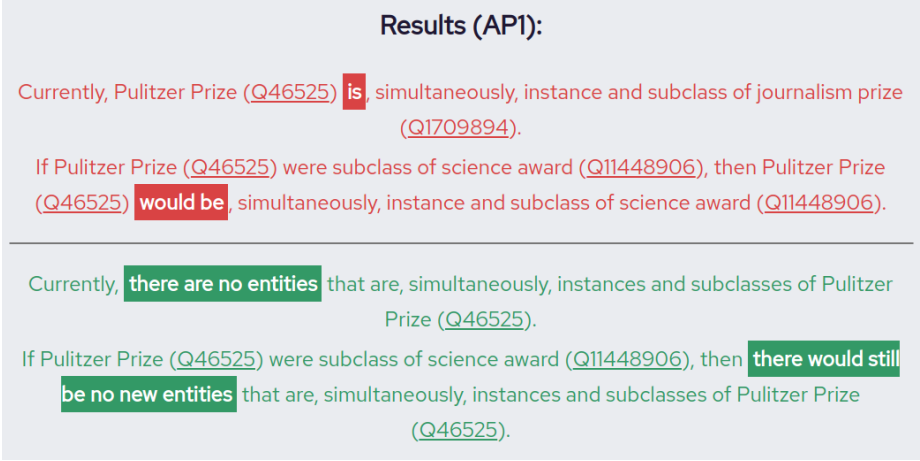

Fig. 5. WAPA results regarding hypothetical statement about `Pulitzer Prize (Q46525)`.

6. Final Considerations

In this paper, we conduct an empirical analysis of the Wikidata platform. We do that as a way to demonstrate how recurrent are anti-patterns exemplifying problems related to the representation of types and instances in large multi-level knowledge models. As this empirical data corroborates, this is a widespread problem with thousands and even million of occurrences in Wikidata. We also identify the items in Wikidata appearing in the highest number of occurrences of these anti-patterns. By conducting a conceptual analysis of these cases, we manage to venture an explanation for their occurrence, and propose interpretation solutions that would eliminate them. Finally, we show how these anti-patterns can inform the construction of automated procedures that can proactively detect these anti-patterns before they are introduced in such a knowledge model. In an earlier work, some of us explored the role of a multi-level modeling language (ML2) in detecting the occurrence of the anti-patterns discussed here [13]. Differently from that work, here we proposed a Web application that can be used by Wikidata users to detect the problems in a language-independent manner.

We should note that the concepts of order and the stratification of taxonomies into consistent multi-level structures are concerns present in Wikidata since revisions introduced in mid 2016. To support stratified taxonomies, the platform includes at the top of its specialization hierarchy a set of classes representing different orders, namely `first-order class (Q104086571)`, `second-order class (Q24017414)`, `third-order class (Q24017465)`, `fourth-order class (Q24027474)`, `fifth-order class (Q24027515)`, and `fixed order metaclass of higher order (Q24027526)`. These classes are declared as equivalent to their counterparts in the OpenCyc ontology [4]. However, they are underused in the platform, and, as we show here and in [13], their mere inclusion in the platform without adequate computational aid is insufficient to prevent the introduction of anti-patterns in new revisions. This motivated us to provide some automated support as shown in this paper.

The dual facet of entities that are both types and instances is a phenomenon that is well-documented in (multi-level) conceptual modeling [6], in formal ontology [14], and in linguistics [15]. In particular, the phenomenon of *systematic polysemy* in language accounts for many cases of this problem. For example, when we say “these ducks in the backyard are common around Europe”, we are making a polysemic reference that overloads the term duck with particular duck instances (those in the backyard) with a duck type (that which is repeatable in a population of ducks and, hence, which is common around Europe). This polysemy that is present in natural language, we conjecture, is also manifested in the construction of lightweight representation structures such as Wikidata. This is specially the case when such a structure is collectively constructed in an asynchronous manner by millions of users, many of which are not expert modelers. This is made worse when these naive modeling strategies (oblivious to these problems) are codified in computer programs (e.g., softbots) that automatically transfer knowledge snippets from other existing data sources.

As we show here, by conducting an analysis of the logical and ontological reasons behind the phenomena causing these semantic confusions, we can proactive devise methodological (e.g., anti-patterns) and computational tools that can assist users in avoiding these mistakes. In this sense, the work presented here is in line with a number of successful initiatives of employing ontological principles to evaluate and rectify large-scale knowledge structures. These include, for example: (i) [16] and [17], which respectively use the DOLCE foundational ontology and the OntoClean methodology for analyzing and proposing correction to the Wordnet Top-level; (ii) [18], which uses a lightweight version of DOLCE (termed DOLCE-Zero) for detecting anti-patterns in DBPedia. The works in (i) focus on detecting taxonomic problems related to ontological notions such as identity, unity and dependence. In contrast, in (ii), the most common patterns detected are related to logical conflicts between disjoint types that are expected by and asserted to given properties. These are related to confusions between objects and events, agents and places, physical and social objects, etc. For example, `dbpedia#AlfonsoXIIofSpain dbo#birthPlace dbpedia#Madrid`, where `dbpedia#Madrid` is erroneously typed as `dbo#Agent` (as a geopolitical entity), which is a confusion between the disjoint types Place and Agent. In (ii), however, one of the patterns detected is what the authors call *metonymy*, which is a conflict arising from disjoint but related interpretations of the same concept. In particular, they make the example of `dbo#family`, which is used to related instances of `dbo#Species` and its property specializing concepts. However, `dbo#Species` are aligned to the type Organism, because "species in DBpedia include species as well as individual exemplars of a species (for example, famous race horses)". Although this case seems to exemplify a type-/instance confusion, the authors arrive at it by, once more, detecting disjoint types in the domain/range of properties, as opposed to explicitly identifying anti-patterns related to this problem. Since the disjointness constraint between individuals and types is entailed by strict stratification (i.e., individuals necessarily belong to a ground strata of uninstantiated entities while types, by definition, are entities that have instances at a lower strata), in our approach, these cases are systematically detected as a particular configuration satisfying one of our anti-patterns (AP1). Moreover, they seem to have a somewhat lenient approach with respect to these problems: "[t]he metonymy anti-pattern is difficult to resolve, because it is due to ambiguities that seem widespread in human language. Metonymy seems related to human propensity for an economy of means... [we try] to accommodate this 'power of ambiguity"'. We here take a radically different approach in this respect by advocating that these problems can cause logical contradictions and conceptual confusion, and by proposing concrete means to detect and correct them.

The analysis conducted in Section 4 was limited to a subset of the top-ranking notions appearing there. In particular, we restricted ourselves to cases of AP2 (Table 2) that were also cases of AP1 (Table 1). In Table 2, however, there are a number of examples that hide subtle ontological and semantic aspects and, hence, that deserve further conceptual analysis. Examples include `information`, `binary relation`, and `class`. These will be addressed in our future work. Finally, this work is part of a long-term effort to monitor the quality and evolution of multi-level taxonomies in Wikidata. We have established a goal to re-assess the state of the platform in 4 year cycles. In the next update cycle, we should be able to observe also evolution trends tracing back to the first assessment in 2016.

References

[1] L. Wetzel, Types and Tokens, in: *The Stanford Encyclopedia of Philosophy*, Fall 2018 edn, E.N. Zalta, ed., Metaphysics Research Lab, Stanford University, 2018.

[2] F. Brasileiro, J.P.A. Almeida, V.A. Carvalho and G. Guizzardi, Applying a Multi-Level Modeling Theory to Assess Taxonomic Hierarchies in Wikidata, in: *Proc. 25th International Conference Companion on World Wide Web*, WWW '16 Companion, 2016, pp. 975–980. ISBN 978-1-4503-4144-8.

[3] A.A. Dadalto, J.P.A. Almeida, C.M. Fonseca and G. Guizzardi, Type or Individual? Evidence of Large-Scale Conceptual Disarray in Wikidata, in: *40th International Conference on Conceptual Modeling (ER 2021)*, Lecture Notes in Computer Science, Vol. 13011, Springer, 2021, pp. 367–377. doi:10.1007/978-3-030-89022-3_29.

[4] D. Foxvog, Instances of instances modeled via higher-order classes, in: *Workshop on Foundational Aspects of Ontologies (FOnt 2005), 28th German Conference on Artificial Intelligence*, 2005, pp. 46–54.

[5] Wikidata, Help:Items — Wikidata, 2021, [Online: 2-May-2021].

[6] V.A. Carvalho and J.P.A. Almeida, Toward a well-founded theory for multi-level conceptual modeling, *Software & Systems Modeling* **17**(1) (2018), 205–231.

[7] P. Gardenfors, Conceptual spaces as a framework for knowledge representation, *Mind and Matter* **2**(2) (2004), 9–27.

[8] J.P.A. Almeida, C.M. Fonseca and V.A. Carvalho, A Comprehensive Formal Theory for Multi-level Conceptual Modeling, in: *Conceptual Modeling*, Springer, 2017, pp. 280–294. ISBN 978-3-319-69904-2.

[9] C. Atkinson, R. Gerbig and T. Kühne, Comparing multi-level modeling approaches, in: *Proc. Workshop on Multi-Level Modelling co-located with ACM/IEEE 17th International Conf. Model Driven Engineering Languages & Systems (MoDELS 2014)*, CEUR Workshop Proceedings, Vol. 1286, CEUR-WS.org, 2014, pp. 53–61. http://ceur-ws.org/Vol-1286/p6.pdf.
[10] L. Wetzel, *Types and tokens: on abstract objects*, MIT Press, Cambridge, Mass, 2009. ISBN 9780262013017.
[11] F.J. Pelletier, On some proposals for the semantics of mass nouns, *Journal of Philosophical Logic* **3**(1) (1974), 87–108.
[12] M. Steen, The Metaphysics of Mass Expressions, in: *The Stanford Encyclopedia of Philosophy*, Winter 2016 edn, E.N. Zalta, ed., Metaphysics Research Lab, Stanford Univ., 2016.
[13] C.M. Fonseca, J.P.A. Almeida, G. Guizzardi and V.A. Carvalho, Multi-level conceptual modeling: Theory, language and application, *Data & Knowledge Engineering* **134** (2021), 101894. doi:10.1016/j.datak.2021.101894.
[14] G. Guizzardi, J.P.A. Almeida, N. Guarino and V.A. de Carvalho, Towards an ontological analysis of powertypes, in: *JOWO@ IJCAI*, 2015.
[15] Y. Ravin and C. Leacock, *Polysemy: Theoretical and computational approaches*, OUP, 2000.
[16] A. Gangemi, N. Guarino, C. Masolo and A. Oltramari, Sweetening wordnet with dolce, *AI magazine* **24**(3) (2003), 13–13.
[17] A. Gangemi, N. Guarino and A. Oltramari, Conceptual analysis of lexical taxonomies: The case of WordNet top-level, in: *Proc. FOIS 2001*, 2001, pp. 285–296.
[18] H. Paulheim and A. Gangemi, Serving DBpedia with DOLCE–more than just adding a cherry on top, in: *International Semantic Web Conference*, Springer, 2015, pp. 180–196.