Semantic Web -1 (2024) 1–18
DOI 10.3233/SW-243562
IOS Press

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] and Giancarlo Guizzardi [b]

[a] *Ontology and Conceptual Modeling Research Group (NEMO), Federal University of Espírito Santo, Brazil*
E-mails: atiliodadalto@protonmail.com, jpalmeida@ieee.org

[b] *Semantics, Cybersecurity & Services (SCS), University of Twente, The Netherlands*
E-mails: c.moraisfonseca@utwente.nl, g.guizzardi@utwente.nl

Editors: Lucie-Aimée Kaffee, University of Southampton, United Kingdom; Simon Razniewski, Max Planck Institute for Informatics, Germany; Pavlos Vougiouklis, Huawei Technologies, United Kingdom
Solicited reviews: Masaharu Yoshioka, Hokkaido University, Japan; Filip Ilievski, Vrije Universiteit Amsterdam, The Netherlands; Two anonymous reviewers

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 instance 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 identifying some of these mistakes.

Keywords: Wikidata, multi-level taxonomies, quality assessment

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" [37]. 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* (numerical identity)?

*Corresponding author: João Paulo A. Almeida, Av. Fernando Ferrari, 514, Vitória, ES, Brazil. E-mail: jpalmeida@ieee.org.

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. 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. It is an instance of car (and *not* a type of car) that can be registered by an owner, be assigned a license plate number, etc.

Logic-based knowledge representation languages such as RDF (and by extension OWL) introduce a special relation to denote instantiation: `rdf:type`, which can be applied to relate an entity to a class it instantiates [34]. In this way, statements such as "Earth is a planet" can be encoded by a corresponding triple constituted by "Earth", `rdf:type` and (the class) "planet".

Despite the presence of a specialized relation to denote instantiation, we can observe in practice difficulties in its use, especially in contrast with the relation of subclassing or specialization which holds between types or classes. Here, there is further ambiguity that arises from the use of 'is a' to express instantiation and subclassing in different contexts. We use quite similar linguistic constructions to say that "the Earth **is a** planet" (denoting instantiation), and to say that "a terrestrial planet **is a** planet" (denoting subclassing). This was recognized early on in the analysis of taxonomic links in semantic networks and discussed in-depth in [4].

This paper examines representation issues that arise from difficulties of navigating these distinctions in practice, by employing Wikidata as a source of empirical data. Wikidata is structured as a graph with millions of nodes called *items*. Wikidata items are used to represent types (classes) (e.g., the item for `planet (Q634)`) or particular individuals (e.g., the item for `Earth (Q2)`). The edges of this graph represent relations between items including specialization and instantiation. When employed correctly, these relations can be used to establish rich *multi-level* taxonomies, when meta-types (such as `astronomical object type (Q17444909)`) are represented in tandem with (first-order) types that instantiate them (such as `planet` and `star (Q523)`) and individuals (such as `Earth` and the `Sun (Q525)`). These taxonomies have been found useful in several domains, including product types [23], biological taxonomy [7,22], organizational structure [6] and software development [19].

Despite the usefulness of multi-level taxonomies, some of us have observed in a first study in 2016 [5] that a large proportion of these taxonomies in Wikidata suffered from quality issues. That study reported the presence of 17,819 classes in multi-level taxonomies in Wikidata, out of which 87.5% were flagged for classification problems, with declarations of instantiation and subclassing making their interpretation ambiguous or leading to inconsistencies. Since then, some other studies revealed quality problems involving the use of instantiation and subclassing, including [26] who corroborated the findings of [5] and identified a large number of "items erroneously treated as classes", and more recently [30] who identified a large number of removed statements over time fixing subclassing and/or instantiation declarations, again corroborating our observations on the difficulties users face when deciding which relation to use. We show in this paper that, several years later, this is a persistent problem, despite certain efforts to address the issues. These include attempts to introduce in Wikidata some metaclasses originally present in OpenCyc [14] and some functionality of "constraint reports".[1]

In multi-level taxonomies in Wikidata, the issues can be observed to occur not only in navigating the distinction between individuals and (first-order) types, but also in distinguishing types of adjacent successive orders (e.g., first-order types and second-order types). The problems were characterized in terms of a number of anti-patterns [5], 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. The continued prevalence of the problems in Wikidata's multi-level taxonomies suggests that further guidelines and tools are required to mitigate the large-scale conceptual disarray.

After eliciting empirical data, we 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

[1] https://www.wikidata.org/wiki/Special:ConstraintReport

them. Finally, we demonstrate a simple automated procedure that can proactively detect these violations before they are introduced to Wikidata. This paper is an extended version of a short paper published in [10]. The short paper only covered one of the anti-patterns we cover here. Further, this paper provides an analysis of the use of the multi-level mechanisms from OpenCyc [14] in Wikidata. Overall, there is more in-depth discussion on the multi-level problems uncovered and better treatment of related work.

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 [5]. 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. Section 6 discusses related work. Finally, Section 7 presents final considerations.

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*" [38]. 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)`, in contrast, 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 Fig. 1). The work of [7] 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.

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.

The domain of biological taxonomy in Wikidata presents a further example of properly stratified multi-level taxonomy. In this domain, an `organism (Q7239)` can be classified by first-order types (such as, e.g., `animal (Q729)`,

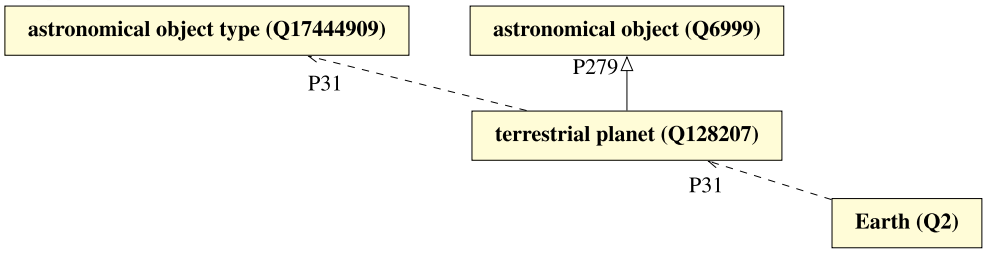

Fig. 1. Terrestrial planet: instance of astronomical object type, subclass of astronomical object.

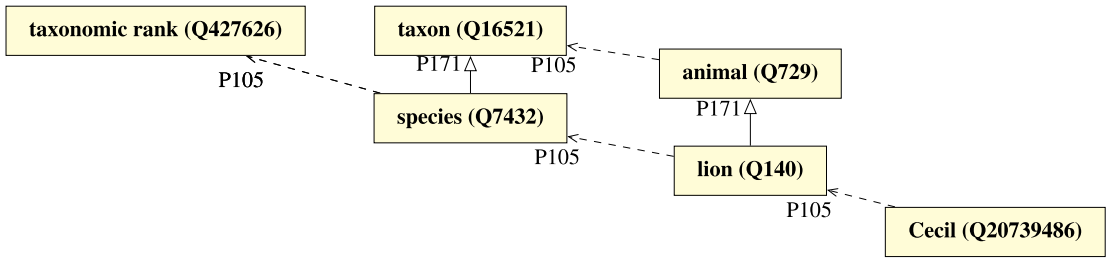

Fig. 2. Four-levels of classification in the biological domain in Wikidata.

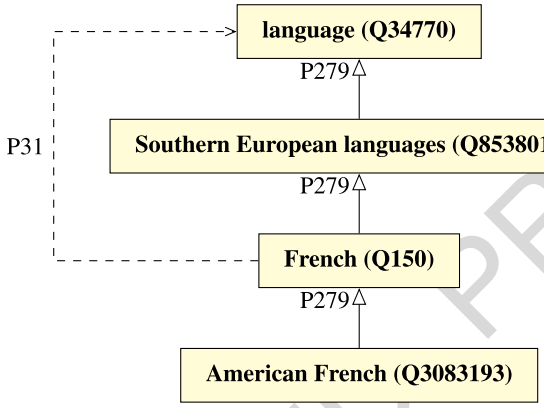

Fig. 3. French as instance and subclass of language.

carnivora (Q25306), lion (Q140)) which are in turn instances of the second-order class taxon (Q16521) and its specializations (e.g., kingdom (Q36732), order (Q36602), species (Q7432)) which are instances of taxonomic rank (Q427626). This example reveals that long chains of instantiation can be established meaningfully: Cecil (Q20739486), a particular lion that lived in Zimbabwe and was killed by a trophy hunter in 2015, is declared to be an instance of lion (a first-order class), which is an instance of species (a second-order class), which is an instance of taxonomic rank (Q427626) (a third-order class). Figure 2 shows some of these items and their relations (taxon rank (P105) and parent taxon (P171) are specialized subproperties of P31 and P279 respectively).

While some multi-level taxonomies structures are clearly structured into individuals, their types, metatypes and so on, this same clear stratification is not present in other taxonomic structures of Wikidata. Consider, for instance, the following fragment concerning the French language, depicted in Fig. 3. 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. At the time of writing in May, 2023, there are 7,062 items that are simultaneously *instances of* and direct or indirect *subclasses of* language.[2] Some of these are clearly classes of languages, such as fictional language (Q2623733), which should not be declared as an instance of language. If treated as an instance of language, properties that are attributed of particular languages, such as writing system (P282), could be incorrectly attributed to

[2]The query and its results are available at https://w.wiki/6ma3.

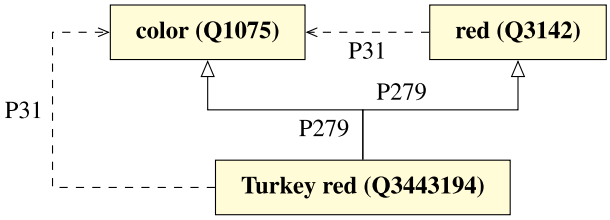

Fig. 4. Turkey red as a specialization of color and of its instance red.

`fictional language` itself. Note that it is instances of `fictional language` like `Klingon (Q10134)` which indeed have a `writing system`, in this case, the fictional alphabet `pIqaD (Q56627865)`).

Similar difficulties are found in the representation of colors, as shown in Fig. 4. 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 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) [18]. (At the time of writing in May, 2023, there are 448 items that are simultaneously *instances of* and direct or indirect *subclasses of* color. There are also 1,803 items that are simultaneously a subclass of color and of one of its instances.[3])

3. Assessment of taxonomic structures in Wikidata

The problems identified in the previous section (and illustrated in Figs 3 and 4) are representative of a large-scale phenomenon involving instantiation and specialization in multi-level taxonomies that was identified originally in [5]. It concerns statements that, when taken together, prevent the stratification of the model into strict metamodeling levels [3] (or orders [7]) consisting of individuals, types, metatypes, metametatypes and so on. Here we set out to revisit this phenomenon. Our objective is to study the occurrence of classification problems in taxonomies in Wikidata as revealed by declarations of instantiation and subclassing, in order to answer the following research questions:

1. How prevalent are the difficulties in establishing the classification order of an item in Wikidata?
2. In which knowledge domains these difficulties occur more frequently?
3. What logical, ontological and semantic explanations could account for these difficulties?
4. Could classification problems be mitigated automatically before they are introduced in Wikidata?

Considering that Wikidata is an open and collaborative platform, it is natural that its continuous evaluation is necessary. There have been a number of developments since the seminal work in [5], including its own impact in the WikiProject Ontology.[4] Also as shown in [30], a large number of changes from `instance of` to `subclass of` declarations occurred over time. Finally, over the years, a number of high-order classes from the OpenCyc ontology [14] were introduced in Wikidata, establishing a basis for declaring classification levels. Hence, we are also set out to understand the impact of these OpenCyc additions in addressing the representational difficulties.

[3]The queries and their results are available at https://w.wiki/6ma9 and https://w.wiki/6maE.

[4]See https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology/Problems#Anti-patterns_from_Multi-Level_Modeling_Theory and https://www.wikidata.org/wiki/Wikidata:Project_chat/Archive/2016/04#Instance_of and https://www.wikidata.org/wiki/Help:Basic_membership_properties#instance_of_(P31)_vs._subclass_of_(P279)_vs._part_of_(P361).

```
1  SELECT DISTINCT ?x ?y WHERE {
2    # the items are related by subclass of (P279) paths (in any direction)
3    # hence, x and y would be at the same order in a stratified scheme
4
5    ?x  ( wdt:P279 | ^wdt:279 )+ ?y.
6
7    # a path relating them includes at least one instance of (P31)
8    # hence x must be at a lower order than y in a stratified scheme
9
10   ?x  ( wdt:P279 | ^wdt:279 | wdt:P31 )* / wdt:P31 /
11                        ( wdt:P279 | ^wdt:279 | wdt:P31 )* ?y.
12 }
```

Listing 1. SPARQL query for fragments involving subclassing and instantiation that defy stratification

3.1. Level stratification

Declarations of `instance of (P31)` are instrumental in defining the classification of items in Wikidata. In a stratified scheme, a declaration of instantiation places the related items in adjacent levels. For example, by stating that `Cecil` is an `instance of lion`, we can infer `Cecil` is at a level lower than `lion`. By stating that `lion` is an `instance of species`, we can infer `lion` is at a level lower than `species`, and so on.

In contrast to instantiation, in a stratified scheme, `subclass of (P279)` declarations place related classes in the same level. To see why this is the case, consider first that, as discussed above, instances in a stratified scheme are placed in one level lower than their classes. Second, in virtue of the semantics of subclassing, the instances of a subclass are also instances of the superclass, and hence must also be placed at one level lower than the superclass (and the subclass). This places both subclass and superclass are at the same level (one level higher than their instances). (For the underlying theory with further details, see [1,7].)

Given these observations concerning instantiation and subclassing in a stratified scheme, we can identify structural patterns (in fact, *anti-patterns*) which indicate fragments that defy stratification a (multi-level) taxonomy when subclassing and instantiation are used in tandem.

We can infer that two items are at different orders when (i) they are directly related by `instance of`; or, when (ii) they are related by complex paths involving `subclassof` segments (in any direction) and at least one `instance of` segment. When two items that we inferred to be at different orders are declared at the same time to be related by `subclassof` paths (in any direction), we have a contradiction. These general conditions for contradiction involving `subclassof` and `instance of` paths are formalized in the SPARQL query in Listing 1, employing complex property paths expressions [35] (the vertical bar | indicates alternative path elements, the caret symbol ^ indicates an inverse path element, i.e., from object to subject, the slash / indicates sequencing; repetitions of any length are denoted with asterisk *, and those denoted with + require at least one occurrence of a path element.)

Given the size of the Wikidata knowledge graph, and the variety of ways in which the paths can be satisfied, checking these conditions for stratification in general is unfeasible. Hence, the approach adopted in [5] was to survey the graph for some specific (simpler) patterns that satisfy these conditions and which could be assessed efficiently at the scale of Wikidata. While this does not guarantee the complete scrutiny of stratification difficulties, it reveals a large number of problematic cases, as we shall see in the sequel. Further, the anti-patterns form clear conceptual structures which can serve as the basis for our conceptual analysis in Section 4.

We retain here the identification of anti-patterns from [5]. The fragment exemplified by the French language is called 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. Listing 2 shows the SPARQL query to find AP1 occurrences considering transitivity for `subclass of` statements. To see why this is a special case of the general pattern presented in Listing 1, note that line 2 in Listing 2 is a special case of line 5 in Listing 1, and line 3 in Listing 2 is a special case of lines 10–11 in Listing 1 (direct instantiation).

```
1 SELECT DISTINCT ?subject ?class WHERE {
2   ?subject wdt:P279+ ?class .
3   ?subject wdt:P31 ?class .
4 }
```

Listing 2. SPARQL query for AP1 (item is simultaneously subclass of and instance of another item)

```
1 SELECT DISTINCT ?subject ?class1 ?class2 WHERE {
2   ?subject wdt:P279+ ?class1 .
3   ?subject wdt:P279+ ?class2 .
4   ?class2 wdt:P31 ?class1 .
5 }
```

Listing 3. SPARQL query for AP2 (item is subclass of two items, one of which is instance of the other)

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. Listing 3 shows the SPARQL query for AP2 with transitivity for subclassing. To see why this is a special case of the general pattern presented in Listing 1, note that lines 2–3 in Listing 3 form a special case of line 5 in Listing 1, also identifying an intermediary item in the path connecting `?class2` (that corresponds to `?x`) and `?class1` (that corresponds to `?y`). Further, line 4 in Listing 3 is a special case of lines 10–11 in Listing 1 (direct instantiation).

3.2. Data collection

In order to deal with the size of Wikidata, we used a filtered dump of the Wikidata database[5] 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[6] 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.

3.3. Anti-pattern occurrences

To assess the occurrence of the anti-patterns, we have executed the SPARQL queries for the anti-patterns in the filtered dump (Listings 2 and 3).

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. 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.

Transitivity of subclassing is important as it reveals a large number of anti-pattern occurrences, which could indicate that it is harder for users to identify the specialization paths to indirect superclasses. The AP1 query without subclassing transitivity (P279) yields 1,279,629 results (42.3% of the 3,022,698 direct `subclass of` statements), 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.

[5] https://zenodo.org/record/4046102

[6] Further dump details and mirrors at https://wdumps.toolforge.org/dump/749.

Table 1
AP1 occurrence ranking – classes that are most frequently instantiated and specialized simultaneously by another item

Place	Wikidata QID	English label	AP1 occurrences	Proportion of instances	Proportion of subclasses
1	Q7187	Gene	971,982	96.8%	94.6%
2	Q8054	Protein	757,360	96.2%	100.0%
3	Q4164871	Position	103,545	91.5%	93.5%
4	Q277338	Pseudogene	49,404	98.9%	98.9%
5	Q427087	Non-coding RNA	49,132	73.9%	99.2%
6	Q2996394	Biological process	30,315	20.1%	100.0%
7	Q12136	Disease	12,293	45.4%	98.6%
8	Q14860489	Molecular function	11,204	99.0%	99.8%
9	Q34770	Language	6,795	61.1%	95.3%
10	Q5058355	Cellular component	4,287	83.2%	99.9%
11	Q294414	Public office	2,544	2.4%	62.4%
12	Q898273	Protein domain	2,493	99.4%	98.5%
13	Q282	Wine	2,143	63.8%	99.9%
14	Q929833	Rare disease	1,994	17.6%	60.4%
15	Q618779	Award	1,469	23.3%	85.9%
16	Q55788864	Developmental defect during embryogenesis	1,403	69.3%	48.7%
17	Q201448	Transfer RNA	1,153	13.2%	99.7%
18	Q11173	Chemical compound	875	0.1%	68.9%
19	Q60754876	Grade of an order	772	95.4%	72.7%
20	Q55789477	Head and neck disease	735	51.4%	85.6%

3.4. *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 (and bound to the `?class` variable in Listing 2) are shown in Table 1 along with the number of times it participates in the anti-pattern. A comprehensive ranking with top 200 of such classes and all scripts used in this paper are available at https://purl.org/nemo/wapa.

Note that a high proportion of the instances and subclasses of these classes in the ranking are involved in the anti-pattern, showing this is a significant problem not only in absolute, but also in relative terms. In the cases of `gene`, `protein`, `pseudogene`, `molecular function`, `protein domain` and `grade of an order`, over 95% of the instances are also declared (directly or indirectly) as subclasses. In the cases of `protein`, `non-coding RNA`, `biological process`, `molecular function`, `language`, `cellular component`, `wine` and `transfer RNA`, over 95% of the (direct and indirect) subclasses are also declared instances.

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), introducing these violations at scale. Other domains highly present in AP1, include social roles and titles (e.g., `position (Q4164871)`,

Table 2
AP2 ranking – classes that most frequently share subclasses with their own instances

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

`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 (bound to the `?class1` variable in Listing 3). These are classes that share subclasses with their own instances. A large number of entities that appear in the ranking for AP1 also appear here: `biological process (Q2996394)`, `position (Q4164871)`, `disease (Q12136)`, etc.

3.5. Anti-patterns statistics considering the OpenCyc basic scheme

Although Wikidata is in principle 'level-blind' [2], 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 [14]. 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 4, only 178 instances of fixed-order classes are found.

Listing 5 shows the SPARQL query for obtaining AP1 occurrences despite explicit fixed-order declarations using the OpenCyc scheme.

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

Listing 4. SPARQL query for classes that are explicitly declared to be fixed-order classes

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

Listing 5. SPARQL query for AP1 with OpenCyc

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)`.

Tables 3 and 4 show that a number of instances of both rankings are declared explicitly as instances of `second-order class`. A star '*' indicates the declaration has been removed from Wikidata at the time of writing. The history of the `anatomical structure` item is instructive, as it has had many edits since the dump, going back and forth in classification; at the time of writing, it is instance of `anatomical entity class type (Q103812671)` which is declared as an instance of `second-order class`).

The results above show that declaring entities to belong to a fixed order 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 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

Table 3
Ranking of occurrences of entities involved in AP1 violations of OpenCyc's fixed-order hierarchies

Place	Wikidata QID	English label	Instance of a fixed-order metaclass?
1	Q7187	Gene	No
2	Q8054	Protein	**Yes (instance of second-order class)**
3	Q4164871	Position	No
4	Q277338	Pseudogene	No
5	Q427087	Non-coding RNA	No
6	Q2996394	Biological process	No
7	Q12136	Disease	No
8	Q14860489	Molecular function	No
9	Q34770	Language	No
10	Q5058355	Cellular component	No
11	Q294414	Public office	No
12	Q898273	Protein domain	No
13	Q282	Wine	No
14	Q929833	Rare disease	No
15	Q618779	Award	No
16	Q55788864	Developmental defect during embryogenesis	No
17	Q201448	Transfer RNA	No
18	Q11173	Chemical compound	**Yes (instance of second-order class)**
19	Q60754876	Grade of an order	No
20	Q55789477	Head and neck disease	No

Table 4
Ranking of occurrences of entities involved in AP2 denoting violations of OpenCyc's fixed-order hierarchies

Place	Wikidata QID	English label	Instance of a fixed-order metaclass?
1	Q2996394	Biological process	No
2	Q4164871	Position	No
3	Q12737077	Occupation	**Yes (instance of second-order class)**
4	Q12136	Disease	No
5	Q28640	Profession	**Yes (instance of second-order class)**
6	Q14860489	Molecular function	No
7	Q294414	Public office	No
8	Q11862829	Academic discipline	**Yes (instance of second-order class)***
9	Q16889133	Class	No
10	Q1207505	Quality	No
11	Q5058355	Cellular component	No
12	Q2424752	Product	No
13	Q4936952	Anatomical structure	**Yes (instance of second-order class)***
14	Q11028	Information	No
15	Q55788864	Developmental defect during embryogenesis	No
16	Q33104279	Philosophical concept	No
17	Q781413	Mental process	No
18	Q130901	Binary relation	No
19	Q18123741	Infectious disease	No
20	Q1914636	Activity	No

fact many tokens of in each cell of our bodies" [36]. 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 [36]. 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))![7] 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 [7], 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 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

[7] In [5], 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.

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" ([25] *apud* [32]).

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 [14] (or orderless [1]) 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).[8] 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 Fig. 3. 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

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

Fig. 5. WAPA results when checking for violations regarding `mining industry (Q1945600)`.

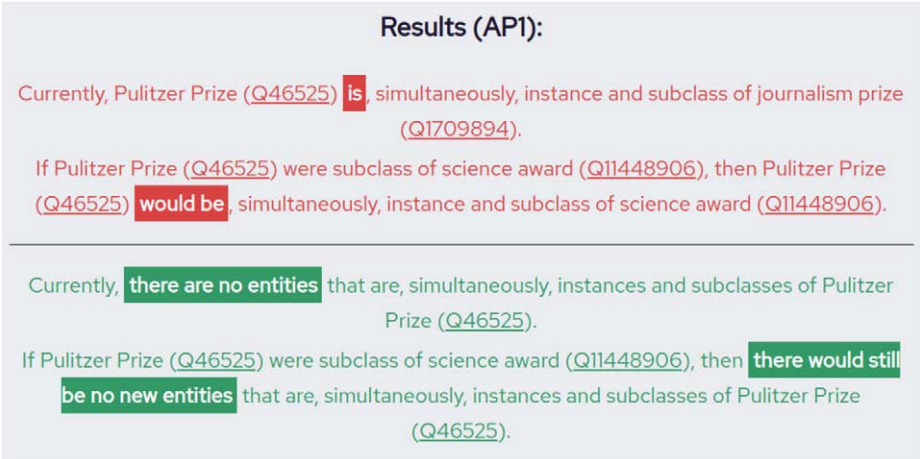

Fig. 6. WAPA results regarding hypothetical statement about `Pulitzer Prize (Q46525)`.

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 Fig. 5, 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.

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 (see Fig. 6).

6. Related work

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) [24], 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. In this respect, [33] raises an important discussion concerning the appropriateness of the Wikidata taxonomy for knowledge representation. He observes that Wikidata reflects the category system of Wikipedia and that it "is more a thesaurus than a classification". In his own words, Wikidata's "practical purpose in many ways is more knowledge organization than knowledge representation". Indeed, this is bound to remain valid if representation issues such as those raised here are not addressed. Even if the consistency requirement is relaxed (and if contradictory statements are allowed in Wikidata), our approach could serve to flag those statements that are mutually inconsistent, possibly qualifying them for further scrutiny.

A number of works have addressed the quality of Wikidata, focusing on different aspects and applying various strategies. A comprehensive literature survey on these aspects was reported in [27]. The authors classified work concerning "Wikidata quality" in several dimensions reminiscent of data quality and discussed for knowledge graphs in general in [12]. These include intrinsic dimensions such as accuracy, trustworthiness and consistency. Our effort here can be understood as addressing consistency, under the assumption of a stratified classification scheme. Also addressing intrinsic dimensions, the authors of [31] point out a number of opportunities for improvement of the taxonomic structures. Their approach was a qualitative one, with a number of illustrative examples, but not explicitly addressing the classification issues we raise. A number of other works address aspects other than intrinsic ones. For example, the authors of [21] report a study on how Wikidata is interlinked to the Linked Open Data (LOD) cloud, while [15] reports on the technical aspects of the rendering of Wikidata content in RDF.

More closely related to the work presented in this paper, [30] dedicates a section to the issue of whether the community of Wikidata users can distinguish classes from instances. They conclude by analyzing switches from `instance of` to `subclass of` (and vice-versa) that these changes are commonly observed. They found 44k cases in which an `instance of` declaration had been replaced with a `subclass of` statement. They also found 444k `subclass of` declarations were replaced by an `instance of` declaration only. They show a small number of problematic cases for illustration purposes, while we aim to characterize the domains in which the problem occurs and their possible ontological remedies. Our approaches are complementary: while they identify changes after they occur, we aim to identify problematic cases to be addressed automatically regardless of community intervention.

Finally, [26] characterized Wikidata quality in its relation to different user roles. They identified two clusters of Wikidata editors: contributors and leaders, with leaders more active than contributors. They concluded that leaders devote more effort to revising properties and taxonomic relations (`instance of` and `subclass of` statements) when compared with contributors. They have found partial support for the hypothesis that higher levels of leader

activity are positively correlated to inheritance richness (number of sub-classes per class), average population (number of instances per class), and average taxonomic depth. Combining the approach taken in the present work with that of [26] could reveal whether leaders contribute to mitigating the classification problems we identify.

Over the years, a number of tools have been developed for Wikidata focusing on different aspects of quality including completeness [11] and vandalism detection [29]. Concerning vandalism detection, the platform includes an AI-based service called Objective Revision Evaluation Service (ORES) that predicts which edits are likely need to be reverted and is integrated into the "Recent changes" panel.[9] Concerning data quality issues, the platform has some support for "constraint reports".[10] These constraints include detection of conflicting statements,[11] and `instance of` and `subclass of` have been declared as such, but only for same classes being specialized (at the time of writing, these are physical object, natural physical object, artificial physical object, musical instrument and equipment). The tool does not consider the transitivity of `subclass of`. As we have shown in Section 3.3, a large number of anti-pattern occurrences (37.1% for AP1) stem from `subclass of` paths of length greater than one. Hence, a simple extension of the functionality as shown for WAPA could already significantly improve problem detection. A further improvement is the detection of problems before they are introduced in Wikidata; given the importance of `subclass of` and `instance of` statements and the difficulties faced by users, the anti-patterns should be checked preemptively during edits.

7. 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 millions 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 [14]. 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 [7,8], in formal ontology [14,20], and in linguistics [28]. In particular, the phenomenon of *systematic polysemy* in language [9] 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

[9] https://en.wikipedia.org/wiki/Special:RecentChanges
[10] https://www.wikidata.org/wiki/Special:ConstraintReport
[11] https://www.wikidata.org/wiki/Q21502838

especially 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 identifying these mistakes. The tool we have implemented does not yet offer suggestions for automated correction. We believe that quick fixes could easily be offered in the form or a choice for the user to remove *subclass of* statements or a seemingly contradictory *instance of* statement in an identified anti-pattern occurrence. But the ultimate correction (which 'quick fix' to apply) requires careful consideration of the semantics of the various items in tandem. Further automating this process is an interesting piece of future work.

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. Since the work discussed here is applicable to any taxonomic system that employs subclassing and instantiation, we intend to investigate the quality of other knowledge graph initiatives from the same perspective. 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.

Acknowledgements

This study was supported in part by CNPq and FAPES.

References

[1] J.P.A. Almeida, C.M. Fonseca and V.A. Carvalho, A comprehensive formal theory for multi-level conceptual modeling, in: *36th International Conference on Conceptual Modeling (ER 2017)*, Lecture Notes in Computer Science, Vol. 10650, Springer, 2017. doi:10.1007/978-3-319-69904-2_23.

[2] 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.

[3] C. Atkinson and T. Kühne, The essence of multilevel metamodeling, in: *International Conference on the Unified Modeling Language*, Springer, 2001, pp. 19–33. doi:10.1007/3-540-45441-1_3.

[4] R.J. Brachman, What IS-A is and isn't: An analysis of taxonomic links in semantic networks, *Computer* **16**(10) (1983), 30–36. doi:10.1109/MC.1983.1654194.

[5] 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, International World Wide Web Conferences Steering Committee*, 2016, pp. 975–980. doi:10.1145/2872518.2891117.

[6] V.A. Carvalho and J.P.A. Almeida, A semantic foundation for organizational structures: A multi-level approach, in: *2015 IEEE 19th International Enterprise Distributed Object Computing Conference*, IEEE Computer Society Press, 2015, pp. 50–59. doi:10.1109/EDOC.2015.18.

[7] 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. doi:10.1007/s10270-016-0538-9.

[8] V.A. Carvalho, J.P.A. Almeida and G. Guizzardi, Using a well-founded multi-level theory to support the analysis and representation of the powertype pattern in conceptual modeling, in: *Proc. 28th Int'l CAiSE Conf.*, Springer, 2016, pp. 309–324. doi:10.1007/978-3-319-39696-5_19.

[9] A. Cruse, *Meaning in Language: An Introduction to Semantics and Pragmatics*, Oxford University Press, Oxford, UK, 2004.

[10] 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.

[11] F. Darari, R.E. Prasojo, S. Razniewski and W. Nutt, COOL-WD: A completeness tool for Wikidata, in: *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks Co-Located with 16th International Semantic Web Conference (ISWC 2017)*, Vienna, Austria, October 23rd-to-25th, 2017, N. Nikitina, D. Song, A. Fokoue and P. Haase, eds, CEUR Workshop Proceedings, Vol. 1963, CEUR-WS.org, 2017, https://ceur-ws.org/Vol-1963/paper466.pdf.
[12] M. Färber, F. Bartscherer, C. Menne and A. Rettinger, Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO, *Semantic Web* **9**(1) (2018), 77–129. doi:10.3233/SW-170275.
[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] 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.
[15] N. Freire and A. Isaac, Technical usability of Wikidata's linked data, in: *Business Information Systems Workshops*, W. Abramowicz and R. Corchuelo, eds, Springer International Publishing, Cham, 2019, pp. 556–567. ISBN 978-3-030-36691-9. doi:10.1007/978-3-030-36691-9_47.
[16] A. Gangemi, N. Guarino, C. Masolo and A. Oltramari, Sweetening WORDNET with DOLCE, *AI magazine* **24**(3) (2003), 13–13. doi:10.1609/aimag.v24i3.1715.
[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. doi:10.1145/505168.505195.
[18] P. Gardenfors, Conceptual spaces as a framework for knowledge representation, *Mind and Matter* **2**(2) (2004), 9–27.
[19] C. Gonzalez-Perez and B. Henderson-Sellers, A powertype-based metamodelling framework, *Software and Systems Modeling* **5**(1) (2006), 72–90. doi:10.1007/s10270-005-0099-9.
[20] G. Guizzardi, J.P.A. Almeida, N. Guarino and V.A. de Carvalho, Towards an ontological analysis of powertypes, in: *Proceedings of the Joint Ontology Workshops 2015 Episode 1: The Argentine Winter of Ontology Co-Located with the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015), Buenos Aires*, Argentina, July 25–27, 2015, CEUR Workshop Proceedings, Vol. 1517, CEUR-WS.org, 2015, https://ceur-ws.org/Vol-1517/JOWO-15_FOfAI_paper_7.pdf.
[21] A. Haller, A. Polleres, D. Dobriy, N. Ferranti and S.J. Rodríguez Méndez, An analysis of links in Wikidata, in: *The Semantic Web*, Springer International Publishing, Cham, 2022, pp. 21–38. doi:10.1007/978-3-031-06981-9_2.
[22] E. Mayr, *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*, Harvard University Press, 1982.
[23] B. Neumayr, K. Grün and M. Schrefl, Multi-level domain modeling with m-objects and m-relationships, in: *Proc. 6th Asia-Pacific Conf. Conceptual Modeling*, Australian Computer Society, Inc., 2009, pp. 107–116. doi:10.5555/1862739.1862754.
[24] 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. doi:10.1007/978-3-319-25007-6_11.
[25] F.J. Pelletier, On some proposals for the semantics of mass nouns, *Journal of Philosophical Logic* **3**(1) (1974), 87–108. doi:10.1007/BF00652072.
[26] A. Piscopo and E. Simperl, Who models the world? Collaborative ontology creation and user roles in Wikidata, *Proc. ACM Hum.-Comput. Interact.* **2**(CSCW) (2018). doi:10.1145/3274410.
[27] A. Piscopo and E. Simperl, What we talk about when we talk about Wikidata quality: A literature survey, in: *Proceedings of the 15th International Symposium on Open Collaboration, OpenSym '19*, Association for Computing Machinery, New York, NY, USA, 2019. doi:10.1145/3306446.3340822.
[28] Y. Ravin and C. Leacock, *Polysemy: Theoretical and Computational Approaches*, Oxford University Press, Oxford, UK, 2000.
[29] A. Sarabadani, A. Halfaker and D. Taraborelli, Building automated vandalism detection tools for Wikidata, in: *Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion*, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, pp. 1647–1654. doi:10.1145/3041021.3053366.
[30] K. Shenoy, F. Ilievski, D. Garijo, D. Schwabe and P. Szekely, A study of the quality of Wikidata, *Journal of Web Semantics* **72** (2022), 100679. doi:10.1016/j.websem.2021.100679.
[31] A. Spitz, V. Dixit, L. Richter, M. Gertz and J. Geiss, in: *State of the Union: A Data Consumer's Perspective on Wikidata and Its Properties for the Classification and Resolution of Entities*, 2021, pp. 88–95. doi:10.1609/icwsm.v10i2.14832.
[32] 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.
[33] J. Voß, Classification of knowledge organization systems with Wikidata, in: *Proceedings of the 15th European Networked Knowledge Organization Systems Workshop (NKOS 2016) Co-Located with the 20th International Conference on Theory and Practice of Digital Libraries 2016 (TPDL 2016)*, Hannover, Germany, September 9, 2016, P. Mayr, D. Tudhope, K. Golub, C. Wartena and E.W.D. Luca, eds, CEUR Workshop Proceedings, Vol. 1676, CEUR-WS.org 2016, pp. 15–22, https://ceur-ws.org/Vol-1676/paper2.pdf.
[34] W3C, RDF 1.2 schema W3C first public working draft, 16 May 2023, 2023, https://www.w3.org/TR/rdf12-schema/.
[35] W3C, SPARQL 1.1 query language, W3C recommendation, 21 March 2013, http://www.w3.org/TR/2013/REC-sparql11-query-20130321/.
[36] L. Wetzel, *Types and Tokens: On Abstract Objects*, MIT Press, Cambridge, Mass, 2009. ISBN 9780262013017.
[37] L. Wetzel, Types and tokens, in: *The Stanford Encyclopedia of Philosophy*, Fall 2018 edn, E.N. Zalta, ed., Metaphysics Research Lab, Stanford University, 2018.
[38] Wikidata, Help: Items – Wikidata, 2021, [online: 2-May-2021], https://web.archive.org/web/20210127110938/https://www.wikidata.org/wiki/Help:Items.