Applying a Multi-Level Modeling Theory to Assess Taxonomic Hierarchies in Wikidata

Freddy Brasileiro[1], João Paulo A. Almeida[1], Victorio A. Carvalho[1,2], Giancarlo Guizzardi[1]
[1] Ontology & Conceptual Modeling Research Group (NEMO), Informatics Department
Federal University of Espírito Santo (UFES), Vitória, ES, Brazil
[2] Research Group in Applied Informatics, Informatics Department,
Federal Institute of Espírito Santo (IFES), Colatina, ES, Brazil
freddybrasileiro@gmail.com, jpalmeida@ieee.org, victorio@ifes.edu.br, gguizzardi@inf.ufes.br

ABSTRACT

Wikidata captures structured data on a number of subject domains, managing, among others, the information underlying Wikipedia and other Wikimedia projects. Wikidata serves as a repository of structured data, whose purpose is to support the consistent sharing and linking of data on the web. To support these purposes, it is key that Wikidata is built on consistent data models and representation schemas, which are constructed and managed in a collaborative platform. In this paper, we address the quality of taxonomic hierarchies in Wikidata. We focus on taxonomic hierarchies with entities at different classification levels (particular individuals, types of individuals, types of types of individuals, etc.). We use an axiomatic theory for multi-level modeling to analyze current Wikidata content, and identify a significant number of problematic classification and taxonomic statements. The problems seem to arise from an inadequate use of instantiation and subclassing in certain Wikidata hierarchies.

CCS Concepts

• Collaborative and social computing systems and tools→Wikis • Knowledge representation and reasoning.

Keywords

Multi-level modeling; metamodeling; Wikidata; taxonomies.

1. INTRODUCTION

The importance of structured data on the web has become clear in the recent years, and has fed developments to make it possible for data to be shared and reused across application, enterprise, and community boundaries [1]. Currently, many initiatives focus on structured data in an effort to facilitate the automated processing of data, as opposed to human consumption through natural language. One such initiative is Wikidata [2], a project of the Wikimedia Foundation to capture the structured data underlying Wikipedia, the popular online encyclopedia, and other Wikimedia sister projects. The content of Wikidata is available under a free license, and can thus be consumed and linked to other data sets on the linked data web[1].

The Wikidata repository consists mostly of *items* and *statements* about these items. *Items* are used "to represent all the things in human knowledge, including topics, concepts, and objects", and are given a unique identifier, a label and a description [3]. *Statements* are used "for recording data about an item", and "consist of (at least) one property-value pair"; they serve to "connect items to each other, resulting in a linked data structure" [4].

In order to organize Wikidata's content, some items (termed "classes") may be used to classify other items through the "instance of" property. For example, the item "London" is related to the item "city" through the "instance of" property, to represent the fact that London is a city. Further, classes can be related through the "subclass of" taxonomic property, defining thus hierarchies of classes, from more general to more specific ones [5]. For example, "city" and "country" are subclasses of "administrative territorial entity", which is a subclass of "human-geographic territorial entity".

In several knowledge domains, classes themselves may be subject to categorization, resulting in classes of classes (or metaclasses). In Wikidata, this means that an item that is a class may itself be related to a (metaclass) through the "instance of" property. For instance, the airplane that was flown solo by Charles Lindbergh on the first non-stop flight from New York to Paris (called "Spirit of St. Louis" with code Q939784 in Wikidata) is "instance of" "fixed-wing aircraft" (Q2875704), which in turn is instance of "aircraft class" (Q20026879). This means that knowledge about this domain includes reference to entities of different (but nonetheless related) classification levels. Other examples of multiple classification levels come from domains such as the biological taxonomy domain [6], software development domain [7] and product types [8].

Given the inherent complexity of dealing with multiple classification levels simultaneously, it is not surprising that the quality of multi-level taxonomic structures in Wikidata is not consistently high. In fact, the conceptual challenges of multi-level classification have given rise to an active area of research which has been referred to as multi-level modeling [8], [9]. Multi-level modeling extends traditional two-level metamodeling with an arbitrary number of metalevels, and the techniques for multi-level conceptual modeling (e.g. [7]–[11]) have focused on providing modeling concepts to deal with types in various classification levels and the relations that may occur between those types.

We argue that the quality of taxonomic structures is key to properly capturing knowledge in Wikidata, and thus, in this paper we leverage the advances in multi-level modeling principles to assess taxonomic hierarchies in Wikidata which employ more than one level of instantiation. We identify issues in a number of hierarchies in Wikidata, as they violate rules of the adopted multi-level theory; the problems seem to arise from an improper use of *instance of* and *subclass of* properties in Wikidata hierarchies. This paper is further structured as follows: section 2 presents the theory for multi-level modeling which is

[1] www.wikidata.org

used in our analysis; section 3 presents cases of multi-level hierarchies in Wikidata, discussing how these are represented and how they can be understood in light of the adopted theory; section 4 presents results of our analysis of current Wikidata content. Finally, section 5 presents some concluding remarks and outlines future investigation.

2. MLT: A THEORY FOR MULTI-LEVEL MODELING

In a recent development, some of us have proposed in [12] an axiomatic theory for multi-level modeling called MLT. The theory is founded on the notion of (ontological) instantiation, which is applied regularly across levels ("orders"). MLT precisely defines a set of structural relations that may occur between elements of different classification levels. It has been used in order to provide foundations for ontology-based conceptual modeling [13] and to analyze the powertype support in UML class diagrams [14]. We present here briefly the fragment of MLT that we use to analyze the hierarchies in Wikidata. For a fuller presentation and complete formal characterization of MLT, the reader should refer to [12].

The notions of types and individuals are central for MLT. According to MLT, types are predicative entities that can possibly be applied to a multitude of entities (including types themselves). Particular entities, which are not types, are considered *individuals*. Each type is characterized by a *principle of application* with which we judge whether the type applies to an entity (e.g., whether something is a Person, a Dog, a Chair) (following [15]). If the principle of application of a type *t* applies to an entity *e* then it is said that *e is an instance of t.*

MLT is formally defined using first-order logic, quantifying over all possible entities (individuals and types). In this formal theory, the *instance of* relation is represented by a binary predicate *iof(e,t)* that holds if an entity *e is instance of* an entity *t* (denoting a type). For instance, the proposition *iof(Vitória,City)* denotes the fact that "Vitória" is an instance of the type "City".

MLT admits types having individuals as instances as well as types that have other types as instances. In order to accommodate these varieties of types, the notion of *type order* is used. Types having individuals as instances are called *first-order types*, types whose instances are first-order types are called *second-order types* and so on.

The axiomatic theory was built up defining the conditions for entities to be considered individuals, using the logic constant "Individual". Thus, *an entity is an instance of "Individual" iff it cannot possibly be related to another entity through instantiation.* The constant "First-Order Type" (or shortly "1stOT") *characterizes the type that applies to all entities whose instances are instances of "Individual".* Each *entity whose possible extension contains exclusively instances of "1stOT" is an instance of "Second-Order Type"* (or shortly "2ndOT"). Analogously *"Third-Order Type"* (or shortly "3rdOT") *characterizes the type that applies to all types whose instances are instances of "2ndOT".* The concept of "Individual" is formally defined in axiom A1 of Table 1. Axioms A2, A3 and A4 (in Table 1) define, respectively, the concepts of "First-Order Type" ("1stOT"), "Second-Order Type" ("2ndOT"), and Third-Order Type ("3rdOT"). We call "Individual", "1stOT","2ndOT", and "3rdOT" the basic types of MLT[2].

Table 1. MLT Rules

A1	$\forall x\ \mathrm{iof}(x, \mathrm{Individual}) \leftrightarrow \nexists y\ \mathrm{iof}(y, x)$
A2	$\forall t\ \mathrm{iof}(t, \mathrm{1stOT}) \leftrightarrow (\exists y\ \mathrm{iof}(y, t) \wedge (\forall x\ \mathrm{iof}(x, t) \rightarrow \mathrm{iof}(x, \mathrm{Individual})))$
A3	$\forall t\ \mathrm{iof}(t, \mathrm{2ndOT}) \leftrightarrow (\exists y\ \mathrm{iof}(y, t) \wedge (\forall t' \mathrm{iof}(t', t) \rightarrow \mathrm{iof}(t', \mathrm{1stOT})))$
A4	$\forall t\ \mathrm{iof}(t, \mathrm{3rdOT}) \leftrightarrow (\exists y\ \mathrm{iof}(y, t) \wedge (\forall t' \mathrm{iof}(t', t) \rightarrow \mathrm{iof}(t', \mathrm{2ndOT})))$
A5	$\forall x \big(\mathrm{iof}(x, \mathrm{Individual}) \vee \mathrm{iof}(x, \mathrm{1stOT}) \vee \mathrm{iof}(x, \mathrm{2ndOT}) \vee \mathrm{iof}(x, \mathrm{3rdOT})\big) \vee (x = \mathrm{3rdOT})$
D1	$\forall t1,t2\ \mathrm{specializes}(t1,t2) \leftrightarrow (\neg \mathrm{iof}(t1,\mathrm{Individual}) \wedge \neg \mathrm{iof}(t2,\mathrm{Individual}) \wedge (\forall e\ \mathrm{iof}(e,t1) \rightarrow \mathrm{iof}(e,t2)))$
D2	$\forall t1,t2\ \mathrm{properSpecializes}(t1,t2) \leftrightarrow (\mathrm{specializes}(t1,t2) \wedge t1 \neq t2)$

It follows from axioms A1 – A4 that: (i) the basic types of MLT have no instances in common i.e., their extensions are disjoint; and (ii) "Individual" is instance of "1stOT" which, in turn, is instance of "2ndOT", which is instance of "3rdOT". Further, according to MLT, every possible entity must be instance of exactly one of its basic types (except the topmost type) (A5 in Table 1). This makes the set of extensions of the basic types a partition of the set of entities considered in the theory (and their union the domain of quantification).

Since the instantiation relation denotes that an element is a member of the extension of a type, it must be irreflexive, antisymmetric and anti-transitive [16]. Further, instantiation relations hold between two elements such that the latter is one order higher than the former. This is a common feature of the *instance of* relation in various techniques which adopt the so-called *strict metamodeling* principle [17]. In our theory, all these properties are guaranteed by axioms A1 – A5.

MLT formally defines the notion of *specialization* between types, as follows: a type (subclass) *t specializes* another type *t'* iff *all instances of t are also instances of t'* (see definition D1 in Table 1). According to this definition every type specializes itself. Since this may be undesired in some contexts, MLT defines the *proper specialization* relation as follows*: t proper specializes t' iff t specializes t' and t is different from t'* (see definition D2 in Table 1). Therefore, while the *specialization* is a reflexive relation, the *proper specialization* relation is irreflexive. Further, the definitions presented thus far guarantee that both *specializations* and *proper specializations* (i) are antisymmetric and transitive, and (ii) may only hold between types of the same order.

The definitions presented so far leads to a basic pattern in MLT: every type that is not one of MLT's basic types (e.g., a domain type) is an instance of one of the basic higher-order types (e.g., "1stOT", "2ndOT" and "3rdOT"), and, at the same time proper specializes the basic type at the immediately lower level. For example, consider a type "Person" that applies to all human beings. Since "Person" applies to individuals (e.g. John or Mary), it is an instance of "1stOT" and proper specializes "Individual". Further, consider a type named "Person Age

[2] For our purposes in this paper, first-, second- and third-order types are enough. However, this scheme can be extended to consider as many orders as necessary [12].

Phase" whose instances are specializations of "Person" (thus, instances of "1stOT") that classify persons according to their age (e.g. "Child" and "Adult"). Thus, "Person Age Phase" is an instance of "2ndOT" and proper specializes "1stOT". Figure 1 illustrates this basic pattern using a notation that is largely inspired in UML. We use the UML class notation to represent both the MLT basic types and the domain types (the theory basic types are shaded to differentiate them from domain elements). Since UML does not allow for the representation of links between classes, we use dashed arrows to represent relations that hold between the types, with labels to denote the names of the predicates that apply. For instance, a dashed arrow labeled *iof* between "Individual" and "1stOT" represents that the former is an instance of the latter (i.e., that *iof(Individual,1sOT)* holds). The traditional UML notation to specializations is used to represent the *proper specialization* relations (e.g. to represent the fact that the proposition *properSpecializes(Person, Individual)* holds). Finally, we use the instance specification notation to represent an individual (e.g. John). For the sake of simplicity we omit the representation of some relations that are implied by the represented relations. For example, although we do not represent that "Adult" *is instance of* "1stOT" it can be inferred by the fact that it *is instance of* "Person Age Phase" which *proper specializes* "1stOT". The notation used to elaborate Figure 1 is used in all further diagrams in this paper.

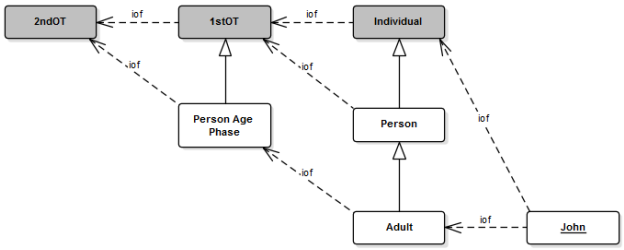

Figure 1. Illustration of MLT basic pattern

Note that the theory results in a model that is stratified according to levels of classification, with specialization only used intra-level, and instantiation used only to related adjacent levels. It is this stratification which will be the main object of our analysis of the content in Wikidata. We explore here the hypothesis that violations of this stratification can allow us to flag potentially inadequate uses of instantiation and subclassing.

3. TAXONOMIC HIERARCHIES IN WIKIDATA

Two properties are central to structure the content in Wikidata: the *instance of* (P31) and the *subclass of* (P279) properties. According to Wikidata, the *instance of* property represents that "*an item is a specific example and a member of another item"* [18]. For instance, considering that *Tim Berners-Lee* is a "specific example" of *Human*, it is stated that *Tim Berners-Lee is instance of Human*. The definition of *instance of* provided in Wikidata is informal and silent about its formal logic properties (symmetry, reflexivity and transitivity). Observing its use in Wikidata content we have concluded that its purpose is similar to the *iof* relation of MLT: to denote that a type applies to an element. Therefore, in order to apply MLT to assess taxonomic hierarchies in Wikidata, we consider the semantics of its *instance of* property to correspond to that of the *iof* relation in MLT.

Wikidata defines *subclass of* as a property that represents that "*all instances of an item are instances of another item"* [19]. For instance, to represent that all instances of *Ship* are also instances of *Watercraft* it is defined that *Ship is subclass of Watercraft.* Further, *subclass of* is characterized as transitive and asymmetric (i.e., antisymmetric and irreflexive). We consider the semantics of the *subclass of* property in Wikidata to correspond to that of the *proper specialization* relation in MLT.

The establishment of the semantics of *instance of* and *subclass of* properties in terms of MLT allow us to use the MLT rules to assess Wikidata content. To illustrate this, we extracted from Wikidata a fragment of a biological taxonomy and the classification of the Cecil lion in such taxonomy. Cecil *is instance of Panthera Leo*, which *is instance of Species.* Species, in its turn, is *instance of Taxonomic Rank.* Considering the definition of *subclass of,* we can conclude that *Cecil* is also *instance of Panthera* and, consequently, of all its super classes. **Figure 2** illustrates this example.. The notational conventions applied in Figure 1 were also used in **Figure 2**. Additionally, in order to increase the readability of the diagram, we use dashed rectangles to group elements that instantiates the same other element and draw only one arrow between the border of the rectangle and the other element.

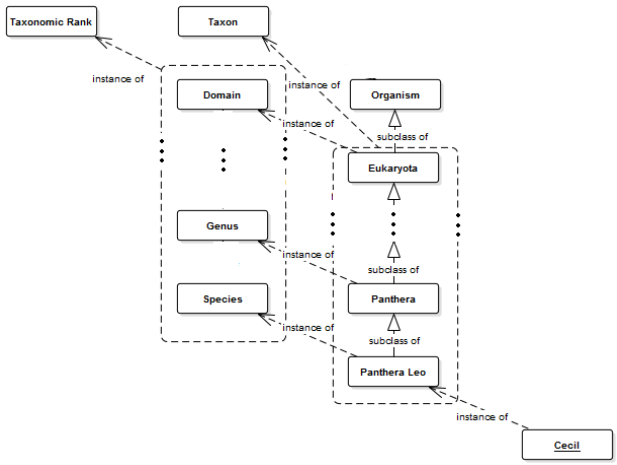

Figure 2. Short representation for Taxonomic Biological Domain in Wikidata

Considering the chain of instantiations in **Figure 2** we can clearly detect a notion of levels: *Cecil*, *Organism*, *Taxon* and *Taxonomic Rank* are at different levels of classification. If we assume *Cecil* as an *instance of Individual*, since we know that it has no instances, we can apply the MLT basic pattern to deduce new information from the diagram in **Figure 2**. First, we can infer that *Panthera Leo* and all its super classes are both *subclasses of Individual* and *instances of 1stOT*. Consequently, the classifiers of *Organism* types (e.g., *Taxon*, *Domain*, *Species*) are both *subclasses of 1stOT* and *instances of 2ndOT*. Finally, *Taxonomic Rank* is inferred as *subclass of 2ndOT* and *instance of 3rdOT.*

The example illustrated in **Figure 2**. conforms to the stratification underlying MLT rules, following its basic pattern. However, there is no automated support or guidelines to prevent a contributor from violating this conformant structure. For example, a modification introducing a second lion (e.g.. "Simba") which is both an instance of *Panthera Leo* and *Species* would go undetected, and would result in an inconsistent hierarchy. In fact, we have observed many occurrences of such potentially problematic hierarchies in current Wikidata content.

For example, take Wikidata information about Tim Berners-Lee and his professional occupation (a fragment of which is depicted in **Figure 3**). Tim is considered *instance of* Computer Scientist. In its turn, Computer Scientist is indirectly *subclass of* Profession. Thus, we can conclude Tim is instance of Profession(!), which clearly violates our sense of what a Profession is. Formally, these statements could be considered inconsistent in the light of MLT: since *instance of* is anti-transitive and Computer Scientist is *instance of* Profession, Tim cannot be *instance of* Profession.

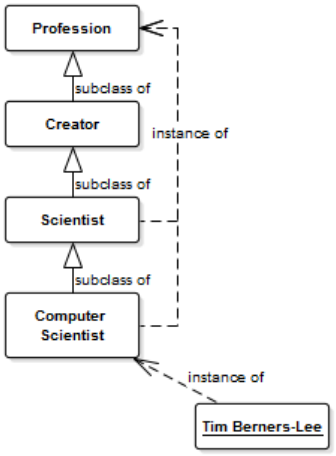

Figure 3. Wikidata information about Tim Berners-Lee and his professional occupation

Now, considering Tim Berners-Lee as Individual, since it has no instances, we can apply the MLT basic pattern to deduce information. First, we conclude that Computer Scientist and all its super classes are both *subclasses of* Individual and *instances of* 1stOT. Consequently, since *instances of* Profession are *instances of* 1stOT, Profession is both *subclass of* 1stOT and *instance of* 2ndOT. Here, we realize that Profession is *instance of* both 1stOT and 2ndOT, which is invalid by A4 (see Table 1).

We have observed similar problems concerning multiple levels of classification in other domains represented in Wikidata, such as transport, software and sports. In section 4, we present the results of some queries we have submitted to Wikidata in order to detect potential problematic scenarios. We highlight some issues identified and discuss them in the light of MLT.

4. ANALYSIS RESULTS

In order to obtain some indication of the use of multi-level hierarchies in Wikidata, we have queried for three simple cases of anti-patterns that violate the aforementioned strict metamodeling principle. To access data, we used the *Simplified and derived RDF dumps* of Wikidata from January 4th, 2016, available at *RDF Exports from Wikidata*[3]. Moreover, we have queried these using SPARQL, where *instance of* and *subclass of* are simplified to *rdf:type* and *rdfs:subClassOf*, respectively. Note that, in this dump, whenever an item is subclass of another item or when it has subclasses or instances, then it is declared to be an instance of *owl:Class* (through the *rdf:type* property).

Figure 4 illustrates the Anti-Pattern 1 (AP1) that looks for pairs of items (A, Z) such that the second one (Z) is simultaneously a *subclass of* A and an *instance of* A. This anti-pattern can appear under many configurations, i.e., subclass (Z) can be a direct *subclass of* A or there may be a chain of *subclass of* properties between the involved items. The fragment illustrated in **Figure 3** (concerning Tim Berners-Lee's professional occupation) includes two occurrences of this anti-pattern with chains of *subclass of* properties of length 2 and 3. Regardless of the size of this chain, the occurrence of this pattern prevents stratification into classification levels, and creates a formal contradiction: classes A and Z would have be simultaneously at the same level (because they are related by specialization) and at adjacent levels (because they are related by instantiation).

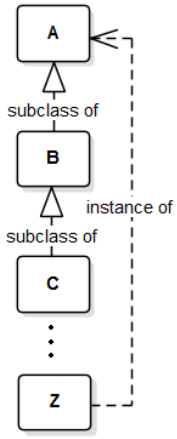

Figure 4. Illustration of Anti-Pattern 1

Table 2 shows the SPARQL query associated to AP1 that considers a transitive closure for *subclass of* statements. For this anti-pattern, we have found in the aforementioned set of Wikidata dumps 14320 occurrences, covering many domains, such as software, sports, biology, food, profession.

Table 2. SPARQL queries for Anti-Pattern 1

```
select distinct * where {
  ?Z rdf:type ?A .
  ?Z rdfs:subClassOf+ ?A .
}
```

Figures 5 and 6 show examples of problematic fragments identified through Anti-Pattern 1. Figure 5 shows that *earthquake* (Q7944) is both *instance of* and *subclass of natural disaster* (Q8065). This fragment seems to have an unclear interpretation. Is an *earthquake* meant to be a *natural disaster* or a special type of *natural disaster*?

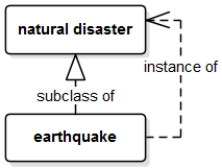

Figure 5. Scenario found in Wikidata for AP1

This lack of clarity that results from the occurrence AP1 has practical implications for the properties of the items involved. For instance, considering that instances of natural disaster are specific *events* (Q1190554), i.e., specific occurrences of *natural disasters* then these instances may be represented as having a *point in time* feature (P585). For example, we can say that the *1985 Mexico City earthquake* took place on September 19th,

[3] http://tools.wmflabs.org/wikidata-exports/rdf/

1985. However, since *earthquake* is also declared to be an *instance of natural disaster* and, thus, an *instance of event*, *earthquake* itself could also be associated to a *point in time*. Notice, however, that *earthquake* is more naturally thought of as a *subclass of natural disaster*, i.e., as a specific *kind of* natural disaster, and a specific kind of event. But, in this case, it would be problematic to attribute a specific *point in time* to this particular class of events. So, in this example, it seems that the undesired relation is the *instance of* relation between earthquake and natural disaster.

Analogously, Figure 6 shows that *Egg waffle* (Q837620) is both *instance of* and indirectly *subclass of food* (Q375). In this case, it is unclear whether an instance of *food*, *waffle* and *Egg waffle* would represent a particular portion of food (the egg waffle John had for breakfast), or a kind of food (such as *waffle* or *Egg waffle*).

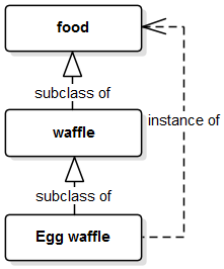

Figure 6. Scenario found in Wikidata for AP1

A second Anti-pattern (AP2) is illustrated in Figure 7. In this case, we have that an item (*C*) has two direct super classes (*A* and *B*) such that one of them is an instance of the other (*B is instance of A*). Similarly to AP1, the occurrences of AP2 present logical inconsistencies that rise from the violation of the strict metamodeling principle. In this case, all *instances of C* are also *instances of A* and *B*. However, *instances of B* cannot be *instances of A* since *B* is itself *instance of A*.

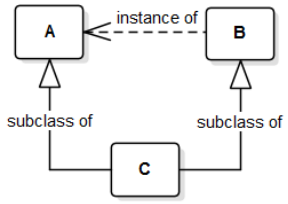

Figure 7. Illustration of Anti-Pattern 2

Table 3 presents a SPARQL query that can be used to detect occurrences of AP2. By running this query, we have found in the aforementioned set of Wikidata dumps 257 occurrences, covering domains, such as diseases, biology, food and colors.

Table 3. SPARQL queries for Anti-Pattern 2

```
select distinct * where {
  ?B rdf:type ?A .
  ?C rdfs:subClassOf ?A .
  ?C rdfs:subClassOf ?B .
}
```

Figure 8 illustrates that *excavator* (Q182661) is an instance of *heavy equipment* (Q874311) and *crawler excavator* (Q5182961) is declared to be a *subclass of* both *excavator* and *heavy equipment*. In another example, *bread* (Q2095) *is an instance of food* (Q7802) and *waffle* (Q375) *is subclass of* both *bread* and *food*. Is *bread* a *food* or a type of *food*? Is *waffle* a *food* or a type of *food*?

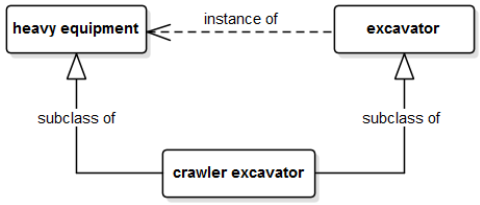

Figure 8. Scenario found in Wikidata for AP2

Finally, a third anti-pattern (AP3) is illustrated in Figure 9. This anti-pattern represents cases in which the anti-transitivity of the instance of relation is violated, making stratification unfeasible. In the case depicted in Figure 9, *C* would have to be simultaneously one and two classification levels below *A*.

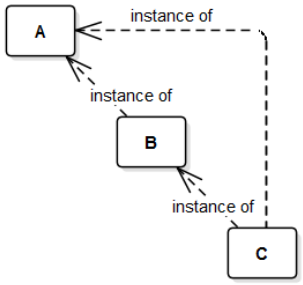

Figure 9. Illustration of Anti-Pattern 3

The query shown in Table 4 can be used to detect instances of AP3. By running it against the aforementioned set of Wikidata dumps, we have found 6708 occurrences of AP3. (Note that since we are concerned only with the *instance of* properties occurring between items in Wikidata, the SPARQL query ignores the triples that declare resources to be *instances of owl:Class;* these are artificial triples introduced in the dump as part of the RDF representation strategy and do not correspond to *instance of* statements in Wikidata.)

Table 4. SPARQL queries for Anti-Pattern 3

```
select distinct * where {
  ?C rdf:type ?B .
  ?B rdf:type ?A .
  ?C rdf:type ?A .
  filter(?A != owl:Class) .
}
```

Figure 10 illustrates *Central Park* (Q160409) as an *instance of* both *urban park* (Q22746) and *park* (Q22698), while *urban park* is also an instance of *park*. This anti-pattern often occurs in chains with terms such as: *award* (Q618779), *Chinese surname* (Q1093580), *family name* (Q101352), *Voivodeship road* (Q1259617), *Mikroregion* (Q11781066), *natural region* (Q1970725).

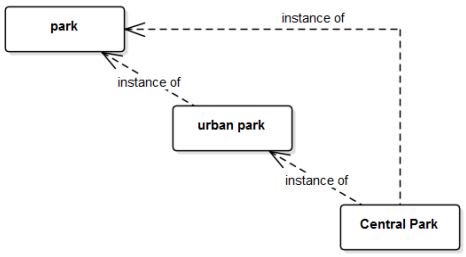

Figure 10. Scenario found in Wikidata for AP3

Table 5 summarizes the results we have obtained from our queries into the Wikidata simplified dump for AP1 and AP2. The total number of classes involved in taxonomic hierarchies is 337102. This number is obtained by counting the items that are either a subject or an object in "subclass of" statements. From this total number of items, 17819 classes are also the object of "instance of" statements, which means they are simultaneously classes and instances of other classes, and thus involved in hierarchies spanning more than one level of classification (our target classes for this investigation). From this number of classes, we have found 15177 classes involved in AP1 (85%) and 441 classes involved in AP2 (2.5%). Thus, a significant percentage of the classes involved in hierarchies spanning more than one level of classification violate the stratification of classification levels.

Table 5. Results for AP1 and AP2

Number of classes in taxonomic hierarchies	337102
Number of classes in taxonomic hierarchies spanning more than one level	17819
Number of classes involved in AP1	15177
Number of classes involved in AP2	441

Table 6 summarizes the results for AP3. Here we contrast: (i) the total number of items in chains of instantiation with three levels (items A, B, and C, where B is an instance of A and C is an instance of B) with: (ii) the number of those items in occurrences of AP3, in which the third item in the chain (C) is also an instance of the first item in the chain (A), violating the stratification. 0,1% of the items that occur in these instantiation chains are found to violate the stratification. The relatively low number of occurrences of this anti-pattern when contrasted with AP1 and AP2 corroborates our intuition that it is the combined use of subclassification and instantiation (a characteristic of ap1 and ap2) that is most challenging to Wikidata contributors.

Table 6. Results for AP3

Number of items in chains of instantiation with three items	6963059
Number of items of AP3 in these chains	7082

5. FINAL CONSIDERATIONS

We have analyzed Wikidata content from the perspective of multi-level modeling. We have observed a number of occurrences of violations of the stratification of levels in Wikidata, which indicate that some support for multi-level modeling could be beneficial in order to support contributors in the collaborative creation of multi-level taxonomies. The queries we have used are the first step in automating this support. In addition to identifying possible problematic occurrences, we understand that more methodological guidance is required for contributors to understand the challenges in multi-level taxonomies and in particular to distinguish clearly between instantiation and specialization.

Future work is required in order to assess whether the items of "class" (Q16889133) and "metaclass" (Q19361238) could be used to provide more explicit support for multi-level modeling in Wikidata. In any case, we have found that these items are rarely employed, and that they seem limited by a three level system (instances, class and metaclass). In the biological taxonomy domain, we see that a fourth level is required (where "Taxonomic Rank" lies). Finally, MLT has a number of other relations to further support structuring multi-level models. These have not been employed here and should be the subject of future works.

6. ACKNOWLEDGMENTS

This research is partly funded by the Brazilian Research Funding Agencies CAPES, CNPq (grants number 311313/2014-0 and 485368/2013-7) and W3C Brasil.

7. REFERENCES

[1] W3C, "W3C Semantic Web Activity." [Online]. Available: https://www.w3.org/2001/sw/. [Accessed: 11-Jan-2016].

[2] D. Vrandečić and M. Krötzsch, "Wikidata: A Free Collaborative Knowledgebase," in *Communications of the ACM*, 2014, vol. 57, no. 10, pp. 78–85.

[3] Wikidata Project, "Help:Items." [Online]. Available: https://www.wikidata.org/wiki/Help:Items. [Accessed: 27-Jan-2016].

[4] Wikidata Project, "Help:Statements." [Online]. Available: https://www.wikidata.org/wiki/Help:Statements. [Accessed: 27-Jan-2016].

[5] Wikidata Project, "Help:Modelling." [Online]. Available: https://www.wikidata.org/wiki/Help:Modelling. [Accessed: 21-Jan-2016].

[6] E. Mayr, *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. 1982.

[7] C. Gonzalez-Perez and B. Henderson-Sellers, "A powertype-based metamodelling framework," *Softw. Syst. Model.*, vol. 5, no. 1, pp. 72–90, 2006.

[8] B. Neumayr, K. Grun, and M. Schrefl, "Multi-level domain modeling with m-objects and m-relationships," in *Sixth Asia-Pacific Conference on Conceptual Modelling (APCCM 2009)*, 2009.

[9] C. Atkinson and T. Kühne, "The Essence of Multilevel Metamodeling," in *4th International Conf. on the Unified Modeling Language*, 2001.

[10] J. Odell, "Power types," *J. Object-Oriented Programing*, vol. 7, no. 2, pp. 8–12, 1994.

[11] L. Cardelli, "Structural subtyping and the notion of power type," *Proc. 15th ACM SIGPLANSIGACT Symp. Princ. Program. Lang. POPL 88*, pp. 70–79, 1988.

[12] V. A. Carvalho and J. P. A. Almeida, "Towards a Well-Founded Theory for Multi-Level Conceptual Modelling," *Int. J. Softw. Syst. Model.*, 2016.

[13] V. A. Carvalho, J. P. A. Almeida, C. M. Fonseca, and G. Guizzardi, "Extending the Foundations of Ontology-based Conceptual Modeling with a Multi-Level Theory," in *34rd International Conference on Conceptual Modeling (ER2015)*, 2015.

[14] 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 *28th International Conference on Advandced Information Systems Engineering (CAISE'16)*, 2016.

[15] G. Guizzardi, *Ontological foundations for structural conceptual models*. Enschede: Telematica Instituut Fundamental Research Series, 2005.

[16] T. Kühne, "Contrasting Classification with Generalisation," in *Proc. of the 6th Asia-Pacific Conference on Conceptual Modeling*, 2009.

[17] C. Atkinson and T. Kühne, "Meta-level Independent Modelling," in *International Workshop on Model Engineering at 14th European Conference on Object-Oriented Programming*, 2000, pp. 1–4.

[18] Wikidata Project, "instance of (P31)." [Online]. Available: https://www.wikidata.org/wiki/Property:P31. [Accessed: 27-Jan-2016].

[19] Wikidata Project, "subclass of (P279)." [Online]. Available: https://www.wikidata.org/wiki/Property:P279. [Accessed: 27-Jan-2016].