

Formal Semantics for the Wikidata Ontology

Peter F. Patel-Schneider, Ege Atacan Doğan

Abstract

A formal model-theoretic semantics for the Wikidata ontology is presented. This semantics is based on simple interpretations very similar to those used in the RDF semantics. Although this semantics does not cover all of Wikidata, notably temporal qualifiers, it can be used to determine an actual meaning for and consequences of the Wikidata ontology, which uses few qualifiers with semantic import.

Wikidata, wikidata.org, is a very large open knowledge graph, containing information on over 122 million entities. Wikidata has a large ontology, with millions of classes, that is embedded in Wikidata like RDFS ontologies are embedded in RDF graphs. Wikidata is open for anyone to edit and the community of Wikidata users is large and varied in expertise. There are also no formal documents providing meaning for the constructs that have been created by the Wikidata community and little to no central authority controlling the creation and use of these constructs. As a result, the meaning of the information in Wikidata can be unclear and different parts of the Wikidata community use constructs in different ways. This is particularly important for the Wikidata ontology, which is meant to control much of the rest of the information in Wikidata.

This paper provides a model-theoretic semantics for many of the constructs that form the Wikidata ontology. Although this semantics ignores some very important parts of Wikidata, the parts that it does ignore do not greatly affect the meaning of the Wikidata ontology. Because there is no central control over Wikidata it is not obvious what part of Wikidata forms the Wikidata ontology. This paper attempts to identify both the constructs that are generally thought by the Wikidata community to form the core of the Wikidata ontology and some other constructs that are part of the Wikidata ontology even though they do not have the same visibility as the core ontology constructs.

This paper does not cover all the constructs that might be considered to be part of the Wikidata ontology. In particular there are several constructs in Wikidata that are related to mereology—the study of parts and wholes—which is often considered to be part of an ontology. These constructs are not fully considered here partly because their meaning in Wikidata is not clear. Future, planned work is indicated to devise a meaning for these constructs that both makes sense and best conforms with the expectations of the Wikidata community.

Even so, this semantics can be used to determine many of the consequences of the Wikidata ontology. This is useful both to show that there are interesting consequences of the Wikidata ontology and to actually show consequences of particular parts of the Wikidata ontology. The hope is that this work will be used to provide a firm, single meaning for the constructs of the Wikidata ontology that will be adopted by the entire Wikidata community.

1. Wikidata

This paper assumes some general level of familiarity with Wikidata and this section only briefly describes the syntax of Wikidata. A longer description of the syntax of Wikidata can be found at www.mediawiki.org/wiki/Wikibase/DataModel.

Wikidata is composed from several primitive syntactic categories. There are items, such as Q42 (Douglas Adams); properties, such as P25 (mother); lexemes; and datatypes, which together form the entities. Items, properties, and lexemes have an opaque identifier with the first letter indicating their

Wikidata'26: Wikidata Workshop at ISWC 2026

EMAIL: pfpschneider@gmail.com (P. F. Patel-Schneider); egeatacandogan@gmail.com (E. A. Doğan)

ORCID: 0009-0000-7789-7459 (P. F. Patel-Schneider); 0009-0000-0854-206 (E. A. Doğan)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

category. Properties also have an associated datatype. The Wikidata datatypes can only be extended by modifying the system implementing Wikidata. The other three categories can be extended at will, although creating new properties requires special privileges and is only done after community discussion.

The treatment of entities here ignores some aspects of entities that are not relevant to the formal semantics for the Wikidata ontology, including labels (except to provide a useful natural language name for the entity), descriptions, aliases, and sitelinks. It also ignores the details required for datatypes, especially the unusual datatypes in Wikidata, as datatypes are not used in the Wikidata ontology.

Wikidata is then conceptually a bag, as opposed to a set, of statements. (Although most presentations of Wikidata show an entity and all the statements that use it as a subject, the statements in an entity are independent.) A *statement* consists of a subject, which must be an item, a property, or a lexeme; a rank of preferred, normal, or deprecated; a main *snak*; a bag of qualifiers, which are *snaks*; and a bag of references, which are also *snaks*. A *snak* consists of a property and a value, which can be an entity or a value belonging to a datatype or the special value *no value* or the special value *unknown value*.

Statements will be written as a sequence of their subject, the property of their main *snak*, the value of their main *snak*, and any qualifiers (property followed by value) in no particular order.

The intended meaning of statements is that a statement is mainly a claim that the subject is related to the value of its main *snak* via the property of its main *snak*. The qualifiers are additional information that can augment the intended meaning of the statement, modify this meaning (called *restrictive qualifiers*, or simply record something about the statement itself. The most important restrictive qualifiers are temporal qualifiers indicating a time period for the claim.

References do not have any effect on this meaning except to provide support for the claim. The initial goal of Wikidata was that all or most statements should have references, which show from where the statement was obtained. Currently, however, many statements in Wikidata do not have references and references are particularly missing from statements that make up the Wikidata ontology.

It also turns out that few¹ of the statements that make up the Wikidata ontology have qualifiers, which the noted exception of statements relating to unions where a qualifier provides the arguments for the union. These qualifiers on unions are, in essence, a form of extended syntax, instead of additional semantic information on a claim.

Wikidata is different from RDF [1] and other knowledge graph formalisms and does not have a dedicated query language. Instead, there is an encoding of Wikidata in RDF and users write SPARQL queries against that encoding, often using the Wikidata Query Service (query.wikidata.org/).

This paper is about formal foundations of the Wikidata ontology. There is no definitive definition of just what constitutes the Wikidata ontology. There are some constructs in Wikidata that are commonly considered to form the core of the Wikidata ontology, permitting creation of a class-based ontology similar in spirit to the ontologies that can be created using RDFS [2]. But there are other constructs in Wikidata that should or could be considered to form part of the Wikidata ontology, including property characteristics, disjointness, class order, and mereology.

So part of the work here involves determining just what constitutes the Wikidata ontology. Making this determination required looking through various pages in the Wikidata wiki, i.e., those pages under wikidata.org that do not form part of Wikidata itself, including discussion pages for Wikidata entities, help pages, and pages created by WikiProject Ontology (www.wikidata.org/wiki/Wikidata:WikiProject_Ontology). Also required was looking at the uses of potential ontology constructs in Wikidata itself, to see both how often these are used and how they are used. A confounding factor in this investigation is that many parts of the Wikidata ontology have been authored or edited by us, or people who have regular communication with us.

¹Of the over 120 million instance statements in Wikidata there were under 250 thousand with restrictive qualifiers; of the over five million subclass statements in Wikidata there were under four thousand with restrictive qualifiers; of the over one thousand subproperty statements in Wikidata there was only one with a restrictive qualifier; all as of 24 July 2026. The restrictive qualifications are almost all temporal.

2. Preliminaries and Related Work

The formal development in this work uses model theory, as opposed to proof theory, because model theory provides a representational grounding in the form of models—simple structures that show the fundamental basis of the work.

The next decision to make when developing a semantics for the Wikidata ontology is whether to formalize the ontology separately from the formalization of the rest of Wikidata, as is generally done in Description Logics [3] and OWL DL [4], or to make the Wikidata ontology part of the regular model theory, as has been done for RDFS [5] and OWL Full [6]. For Wikidata the choice is obvious—the ontology items and properties are just regular items and properties and thus should be treated similarly.

The next decision involves qualifiers in Wikidata. The question is whether to include qualifiers in the model theory, as is done in multi-attributed relational structures (MARS) [7, 8], or to have Wikidata properties be modeled as simple binary relations. The former, which has already been used as a general model theory for Wikidata [9, 10], includes labeled property graph attributes (Wikidata qualifiers) as extra labeled arguments for predicates (Wikidata properties). This results in a more faithful rendering of the syntax of Wikidata, but at the cost of a much more complex model theory and a less intuitive reading of truth. In particular, *restrictive* qualifiers, including especially temporal qualifiers, show up as extra arguments, obscuring their intended meaning.

The approach taken here is to have a simple model theory that reflects the intended meaning of the main snafu of a statement. So a Wikidata property is interpreted as a set of pairs. This ends up with interpretations being very similar to those in the RDF and RDFS semantics [5]. This approach is insufficient to capture the intended meaning of all of Wikidata. Notably it cannot handle temporal qualifiers, which are common in Wikidata. But MARS also has problems with temporal qualifiers and a full formal treatment of Wikidata will be complex. Fortunately there are very few semantically relevant qualifiers on the statements that make up the Wikidata ontology, so this approach is workable, at least to provide a starting point for further work.

So formal meaning in this work is based on simple interpretations. An interpretation, I , is a structure consisting of

1. R , a non-empty set of resources;
2. P , a subset of R , a set of predicates;
3. D , a non-empty set of data values, disjoint from R ;
4. S , a mapping from entities (items, properties, etc.) into R that maps properties to elements of P ;
5. L , a mapping from data literals into D ; and
6. EXT , a mapping from P into subsets of $R \times (R \cup D)$.

To make some formulae simpler, I will be overloaded to also refer to the union of the S and L maps.

Throughout much of this paper there will be an implicit or explicit interpretation and the above symbols will refer to its components without a direct definition of the interpretation in terms of its components.

Our treatment of Wikidata statements ignores deprecated statements and treats preferred and normal statements the same. Deprecated statements are generally considered in Wikidata as not necessarily being true. In the Wikidata ontology preferred statements are generally used to signal the most important type or generalization, which is often just more specific than the others. Our treatment of Wikidata statements also ignores references. As stated above, our treatment of Wikidata statements also generally ignores qualifiers, but see below for the important exception needed for unions.

An interpretation I *satisfies* a non-deprecated statement with subject U , property P , and value *no value* when

$$\forall r \in R : \neg \langle S(U), r \rangle \in EXT(S(P))$$

An interpretation I *satisfies* a non-deprecated statement with subject U , property P , and value *some value* when

$$\exists r \in R : \langle S(U), r \rangle \in EXT(S(P))$$

Normally there would need to be some extra mapping to carry the existential nature of the special *some value* used in Wikidata to indicate an unknown value, but in Wikidata there is only one such special value and it is, in essence, an anonymous variable so there is no need to keep track of its mapping.

An interpretation I satisfies a non-deprecated statement with subject U , property P , and value V otherwise when

$$\langle S(U), I(V) \rangle \in EXT(I(P))$$

An interpretation I satisfies a bag of statements if it satisfies each non-deprecated statement in the bag. Note that there is no difference if the bag is converted to a set by removing all extra copies of statements in the bag. The use of bag as opposed to set is simply because it is possible for there to be multiple copies of a statement in Wikidata.

Then *entailment* is defined in the usual way where a bag of statements, KB , entails a statement U if every interpretation of KB_1 is also an interpretation of U . So (the current contents of) Wikidata entails a statement if all interpretations of the statements in (the current contents of) Wikidata entails the other statement. Both of these can be trivially extended to entailing bags of statements.

3. The Core of the Wikidata Ontology

The goal of this work is to augment the above model theory with semantic constraints that give meaning to the constructs of the Wikidata ontology, starting with its core.

The core of the Wikidata ontology is formed from the constructs that are generally accepted by the Wikidata community and needed to create a class-based ontology with a generalization hierarchy for both classes and properties. Two properties are central to classes, and universally² considered to be part of the Wikidata ontology, namely P31 (instance of)³ and P279 (subclass of). Added to these two properties are the property that supports a generalization relationship for properties, P1647 (subproperty of), and three classes that represent entities, classes, and properties in Wikidata, Q35120 (entity), Q16889133 (class), and Q18616576 (Wikidata property).⁴ This looks very much like the core of RDFS [2] except that there are no domains and ranges. (Wikidata has constraints www.wikidata.org/wiki/Help:Property_constraints_portal that one might take the place of RDFS's domain and range, but Wikidata constraints are only advisory and cannot be considered to have this role.)

The consensus intended meanings for these constructs are very similar to the intended meanings for the RDFS constructs, so their semantic conditions are just simple variations of the RDFS semantics [5]. (Note that all the subsequent definitions and conditions are relative to an interpretation, I .)

A couple of preliminary definitions help with the later semantic conditions, the first defining class extensions in terms of P31 (instance of) and the second defining the set of classes as the class extension of Q16889133 (class):

$$\begin{aligned} CEXT(y) &:= \{x : \langle x, y \rangle \in EXT(S(P31))\} \\ C &:= CEXT(S(Q16889133)) \end{aligned}$$

Then the class extension of Q18616576 (Wikidata property) is required to be the set of properties in the interpretation. (This is different from the definition of classes as the class extension of Q16889133 (class) because P is part of the interpretation.)

$$CEXT(Q18616576) = P \tag{1}$$

²Although the entire Wikidata community appears to consider these two properties to be central to the Wikidata ontology, there is actually considerable misuse of them and some of this misuse appears to be by design.

³Throughout this paper Wikidata entities will be referenced in text using their identifier and their English label and in formulae just using their identifier. Note that labels do not uniquely identify Wikidata entities.

⁴It would seem that this entity should have label “property” to be analogous to Q16889133 (class), but “property” is only an alias for this item.

Next, $P31$ (instance of), $P279$ (subclass of), and $P1647$ (subproperty of) have certain requirements on arguments. This can't be done using domains and ranges, as in RDFS, because Wikidata doesn't have these constructs.

$$\text{if } \langle x, y \rangle \in EXT(P31) \text{ then } y \in C \quad (2)$$

$$\text{if } \langle x, y \rangle \in EXT(P279) \text{ then } x \in C \text{ and } y \in C \quad (3)$$

$$\text{if } \langle x, y \rangle \in EXT(P1647) \text{ then } x \in P \text{ and } y \in P \quad (4)$$

Finally, $P279$ (subclass of) and $P1647$ (subproperty of) are transitive and asymmetric. Note that Wikidata is different from RDFS here, as subclass and subproperty are reflexive in RDFS. In essence, Wikidata uses proper subclass and subproperty, not the permissive versions.

$$I(P279) \text{ is transitive and asymmetric} \quad (5)$$

$$I(P1647) \text{ is transitive and asymmetric} \quad (6)$$

$P279$ (subclass of) has $Q35120$ (entity) as its necessary maximum

$$\text{if } c \in C \text{ then } \langle c, S(Q35120) \rangle \in EXT(S(P279)) \quad (7)$$

Then main effects of $P279$ (subclass of) and $P1647$ (subproperty of) are to create the generalization relationship for classes and properties, respectively.

$$\text{if } \langle x, y \rangle \in EXT(P279) \text{ then } CEXT(x) \subseteq CEXT(y) \quad (8)$$

$$\text{if } \langle x, y \rangle \in EXT(P1647) \text{ then } EXT(x) \subseteq EXT(y) \quad (9)$$

This completes the model theory for the core of the Wikidata ontology. As stated above, it is very similar to the model theory for RDFS.

One interesting consequence of this similarity is that a complexity result for RDFS reasoning carries over to Wikidata, namely that reasoning is inherently serial [11]. This has the further consequence that a SPARQL query cannot be constructed that is guaranteed to always determine class membership in Wikidata. Fortunately, this is only a theoretical result as it would require an expanding intertwining of subclasses and subproperties in a way that is very unlikely to happen. Currently this intertwining does not happen at all. Unfortunately, correctly querying for class membership in current Wikidata is almost never performed even though there are subproperties of $P279$ (subclass of). This is likely because the correct query is hard to write and is slow in the current Wikidata query service.

4. The Near Core

Once one goes beyond the core of the Wikidata ontology it is not obvious just what constructs are considered by the Wikidata community to be part of the Wikidata ontology. There is $Q115490628$ (fundamental Wikidata entity), which is described in English as “property or item in Wikidata that allows logical inferences about other properties, items or values in Wikidata to be made and thus enables reasoning about Wikidata in general”. From the description it seems that all the entities in Wikidata that are important for the Wikidata ontology should be instances of this class.

4.1. Features of Properties

The instances of $Q115490628$ (fundamental Wikidata entity) do include $P31$ (instance of), $P279$ (subclass of), $P1647$ (subproperty of), and $Q16889133$ (class), but not, as of 24 July 2026, $Q35120$ (entity) or $Q18616576$ (Wikidata property). So this class doesn't include even all the concepts identified above as forming the core of the Wikidata ontology.

This class does have as instances several classes and properties that provide features for Wikidata properties, namely Q18647515 (transitive Wikidata property), P6609 (transitive over), P1696 (inverse property), and P11317 (negates property). This indicates that the Wikidata ontology should include some aspects of OWL that are sometimes called RDFS++, although the properties above go beyond what is generally included in RDFS++. Strangely there are some classes that should be in this group that are not instances of this class, namely Q18647518 (symmetric property), Q18647519 (asymmetric Wikidata property), and Q18647521 (reflexive property). There is also Q1097991 (irreflexive relation). This class does not have any properties as instances, but it is easy to include here.

The semantic conditions for these constructs are then:

$$\text{if } p \in CEXT(Q18647515) \text{ then if } \langle x, y \rangle \in EXT(p) \text{ and } \langle y, z \rangle \in EXT(p) \text{ then } \langle x, z \rangle \in EXT(p) \quad (10)$$

$$\text{if } \langle p, q \rangle \in EXT(P6609) \text{ then if } \langle x, y \rangle \in EXT(p) \text{ and } \langle y, z \rangle \in EXT(q) \text{ then } \langle x, z \rangle \in EXT(p) \quad (11)$$

$$\text{if } \langle p, q \rangle \in EXT(P1696) \text{ then } \langle x, y \rangle \in EXT(p) \text{ iff } \langle y, x \rangle \in EXT(q) \quad (12)$$

$$\text{if } \langle p, q \rangle \in EXT(P11317) \text{ then } \langle x, y \rangle \in EXT(p) \text{ iff } \langle x, y \rangle \notin EXT(q) \quad (13)$$

$$\text{if } p \in CEXT(Q18647518) \text{ then } \langle x, y \rangle \in EXT(p) \text{ iff } \langle y, x \rangle \in EXT(p) \quad (14)$$

$$\text{if } p \in CEXT(Q18647519) \text{ then if } \langle x, y \rangle \in EXT(p) \text{ then } \langle y, x \rangle \notin EXT(p) \quad (15)$$

$$\text{if } p \in CEXT(Q18647521) \text{ then } \forall x \in R \langle x, x \rangle \in EXT(p) \quad (16)$$

$$\text{if } p \in CEXT(Q1097991) \text{ then } \forall x \in R \langle x, x \rangle \notin EXT(p) \quad (17)$$

It is unclear why there are not classes like “symmetric Wikidata property”. The above classes that do include “Wikidata property” in their label should be subclasses of Q18616576 (Wikidata property), so every interpretation must satisfy the following axioms:

$$Q18647515 \text{ } P279 \text{ } Q18616576 \quad (18)$$

$$Q18647519 \text{ } P279 \text{ } Q18616576 \quad (19)$$

These statements are present in Wikidata.

It is the case in Wikidata that P279 (subclass of) and P1647 (subproperty of) are instances of Q18647515 (transitive Wikidata property) and Q18647519 (asymmetric Wikidata property), so the semantic conditions 5 and 6 are mirrored in Wikidata itself. (Strangely P279 (subclass of) is related to itself by P6609 (transitive over), but this is the same as being transitive.) It is also the case that P31 (instance of) has P279 (subclass of) as a value for P6609 (transitive over), so the semantic condition 8 is also mirrored in Wikidata itself.

The class fundamental Wikidata entity (Q115490628) also includes Q61719274 (non-restrictive qualifier), Q61719275 (restrictive qualifier), and Q122233685 (restrictive or non-restrictive qualifier) where a restrictive qualifier is one that modifies the meaning of a statement, as opposed to a non-restrictive qualifier that just augments the meaning. It is unclear why non-restrictive qualifier is in this class, but teasing out what is going on here is not necessary at here as this treatment is largely ignoring qualifiers.

The other instances of fundamental Wikidata entity (Q115490628), P10726 (class of non-item property value), P13353 (provides data for property), Q21146257 (type), and Q115449020 (Wikidata property to express the absence of something), do not appear to provide information relevant to the Wikidata ontology. But there are other Wikidata constructs that are relevant to the Wikidata ontology that are not included in this class.

4.2. Unions

Wikidata does not have a property that directly states that two classes are disjoint, which is an important means for finding problems in a knowledge base. What it does have are constructs that define a class as a

union or disjoint union of other classes, mirroring constructs in OWL. The second construct allows the expression of disjointness in Wikidata. The two union constructs use qualifiers in an unusual way.

A union or disjoint union construct is a qualified statement with the union class as subject, P2737 (union of) or P2738 (disjoint union of) as property, and Q23766486 (list of values as qualifiers) as value. The classes being unioned or disjoint unioned are then values of qualifiers with property P11260 (list item).

A complete treatment of qualifiers in Wikidata is far too complex for this paper, so the meaning of union constructs will be done as a special extra condition on these kinds of Wikidata statements.

An interpretation I satisfies a non-deprecated statement with subject U ; property P2737 (union of) or P2738 (disjoint union of); object Q23766486 (list of values as qualifiers); and qualifiers with property P11260 (list item) and values C_i for $0 \leq i \leq n$ when

$$CEXT(S(U)) = CEXT(S(C_1)) \cup \dots \cup CEXT(S(C_n)) \text{ for } 1 \leq i \leq n$$

and

$$CEXT(I(C_i)) \cap CEXT(I(C_j)) = \emptyset \text{ for } 1 \leq i < j \leq n$$

if the property is P2738 (disjoint union of).

Note that this is an open-ended syntax, like the syntax of RDFS containers, where a new qualifier can be added to union or disjoint union. This makes Wikidata non-monotonic, which can be most easily seen by adding a second element to the list. Before the addition, each instance of S is also an instance of C_1 , but after the addition instances of S can be instances of either C_1 or C_2 . Another feature of unions is that a class that is a union or disjoint union with no list elements cannot have any instances, but this feature does not appear to be used in Wikidata.

5. Metaclasses and Class Order

Classes in Wikidata are entities and thus can themselves be instances of classes. This gives rise to metaclasses, classes that have only classes as instances. Wikidata has Q19478619 (metaclass), the class of all metaclasses, which is formalized as the following semantic condition:

$$CEXT(S(Q19478619)) = \{x : \forall y \in CEXT(x) \ y \in C\} \quad (20)$$

Wikidata has two properties that involve metaclasses.

The first, P8225 (is metaclass for), relates a metaclass to the class of the instances of the metaclasses instances. A good example of this in Wikidata is Q2235308 (ship type), whose instances are classes whose instances are ships. So Q2235308 (ship type) has the value Q11446 (ship) for P8225 (is metaclass for).

This relationship is formalized as

$$\text{if } \langle x, y \rangle \in EXT(S(P8225)) \text{ then } \forall z \in CEXT(x) \ CEXT(z) \subseteq CEXT(y) \quad (21)$$

The second property is P2445 (metasubclass of), which, from the English description relates two metaclasses such that the instances of the first are *likely* to be a subclass of an instance of the second. Formalizing this property is problematic. The word “likely” is in the description which indicates that any semantic condition might not be absolute. There are also under 50 uses of the property in Wikidata as of 24 July 2026, many of which involve manufactured goods models, which are themselves problematic in Wikidata, making it hard to determine meaning from actual usage.

Removing the “likely” results in a semantic condition that does seem reasonable, however.

$$\text{if } \langle x, y \rangle \in EXT(S(P2445)) \text{ then } \forall z \in CEXT(x) \ \exists w \in CEXT(y) \ CEXT(z) \subseteq CEXT(w) \quad (22)$$

Both these properties have obvious requirements on their domains and ranges, as follows:

$$\text{if } \langle x, y \rangle \in \text{EXT}(S(P8225)) \text{ then } x \in \text{CEXT}(S(Q19478619)) \text{ and } y \in \text{CEXT}(S(Q16889133)) \quad (23)$$

$$\text{if } \langle x, y \rangle \in \text{EXT}(S(P2445)) \text{ then } x, y \in \text{CEXT}(S(Q19478619)) \quad (24)$$

Class order is an important structuring notion in ontologies with metaclasses. A class is first-order if none of its instances are classes. A class is second-order if all of its instances are first-order classes, and so on. Class order is particularly important in large ontologies like the Wikidata ontology to prevent mixing of different kinds of entities. This may be most important to prevent first-order classes, like Q11446 (ship), from containing classes, in particular instances of Q2235308 (ship type). Class order also induces disjointness, as classes of different orders are disjoint.

Class order can be defined inductively, with non-classes as order 0, as:

$$\begin{aligned} O_0 &:= \{x \in R : x \notin C\} \\ O_n &:= \{x \in R : \forall y \in \text{EXT}(x) \ y \in O_{n-1}\}, \text{ for } n > 0 \end{aligned}$$

Wikidata has a group of classes related to class order, namely Q104086571 (first-order class), Q24017414 (second-order class), Q24017465 (third-order class), Q24027474 (fourth-order class), Q24027515 (fifth-order class), Q24027526 (fixed-order class of higher order), Q23959932 (fixed-order class), and Q23958852 (variable-order class). Each of these classes has a fairly obvious semantic condition.

$$\text{CEXT}(S(Q104086571)) = O_1 \quad (25)$$

$$\text{CEXT}(S(Q24017414)) = O_2 \quad (26)$$

$$\text{CEXT}(S(Q24017465)) = O_3 \quad (27)$$

$$\text{CEXT}(S(Q24027474)) = O_4 \quad (28)$$

$$\text{CEXT}(S(Q24027515)) = O_5 \quad (29)$$

$$\text{CEXT}(S(Q24027526)) = \bigcup_{i>5} O_i \quad (30)$$

$$\text{CEXT}(S(Q23959932)) = \bigcup_{i>0} O_i \quad (31)$$

$$\text{CEXT}(S(Q23958852)) = C \setminus \text{CEXT}(S(Q23959932)) \quad (32)$$

From these semantic conditions it can be deduced that the fixed-order classes are disjoint from fixed-order classes with different orders. In Wikidata it is stated that Q23959932 (fixed-order class) is the obvious disjoint union of the fixed-order classes above. The other obvious disjoint union stated in Wikidata is that Q16889133 (class) is the disjoint union of Q23959932 (fixed-order class) and Q23958852 (variable-order class).

6. Mereology

Mereology is about the relationship between parts and wholes, as in Q1062 (Lake Ontario) is P361 (part of) the Q7347 (Great Lakes). There may be different characteristics of part of, such as being transitive or anti-symmetric, and there may be several varieties of part of, such as physical part. There might also be an inverse property, like P527 (has part). It would seem that formalizing the mereology part of the Wikidata ontology would be easy.

But the mereology properties in Wikidata are used not just to relate individual parts to individual (non-class) wholes, as above, but also to relate classes to other classes, as in Q7362 (ear) is P361 (part

of Q23640 (head). There is also P2670 (part(s) of the class), which appears to relate individuals to classes that provide typing for parts. The intended meaning of these sorts of statements, however, is unclear. As well, there are other properties in Wikidata that appear to be related to mereology, such as P1552 (has characteristic), also with unclear meaning.

If classes in Wikidata were separate from individuals and this determination could be easily made, it would not matter too much that P361 (part of) was overloaded, and it might be possible to tease out the meaning of this overloading. But neither of these are true. Classes are instances of other classes, so P361 (part of) on a class might be stating a part-hole relationship on the class itself or it might be stating something about the classes instances. Worse, there are many problems with class order in Wikidata, resulting in statements entailing (using the semantic constraints above) that actual individuals are incorrectly classes. This means that for many (or even most) parthood statements, it is not possible to determine whether it is about the entity or the entity's instances.

There are some common features of parthood relationships that might be claimed in Wikidata, however. Imagining a scenario that class-class and individual-individual relationships can clearly be separated, the parthood relationship between physical objects (or subclasses of physical objects) would have the asymmetry and transitivity features. These would be expressed through:

P361(*partof*) is a transitive and asymmetric property
P527(*haspart(s)*) is a transitive and asymmetric property

For the more specific P2670 (part(s) of the class), it's possible to create a normalization related to class order, at least for individuals:

$$\text{if } \langle a, b \rangle \in EXT(S(P2670)) \text{ and } b \in O_0 \text{ then if } b \in O_i \text{ then } i = 1 \quad (33)$$

But P2670 (part(s) of the class) is not used very much, as most parthood relationships relating individuals to classes use P361 (part of), so this is a very niche case.

Because of the class order issues, formalizing the current mereology of Wikidata requires prerequisite work in other aspects of the ontology. Because of the unclear meaning parthood relationships for classes the intended meaning of mereological statements also require further investigation and further consultation with the Wikidata community.

7. Conclusion

This paper provides a model theory for the Wikidata ontology that provides meaning and thereby consequences for the Wikidata ontology. A reason for using a simple model theory is to limit the choices that need to (or can) be made.

This is just the start of work in this area. The next step is to determine exactly what the consequences of this model theory are for Wikidata and present them to the Wikidata community so that the community can approve these consequences or provide input on what changes need to be made. Hopefully the changes can be captured without changing the basic nature of the model theory.

There is reason to believe that this will not be a simple process, however. There are large parts of Wikidata where classes have fewer instances than the number of relevant entities in Wikidata, such as Q84048852 (female human). Instead of membership in these classes, female humans are represented as instances of Q5 (human) with P21 (sex or gender) Q6581072 (female). It might be necessary to devise some way of providing the correct meaning for these classes. There are also other places in Wikidata where P31 (instance of) appears to be used in place of or in addition to P279 (subclass of). It might be necessary to devise some principled way of accounting for this apparent conflation.

The investigation in this paper has exposed a problem with fundamental Wikidata entity (Q115490628) missing some constructs that have consequences in the Wikidata ontology. This class should be extended to cover these constructs and maybe some of the non-reasoning constructs in it should be removed. This may have already been done by the time this paper is published.

This work has not investigated mereology. Extending this work to mereology is an obvious next step. However, the meaning of many mereology-related properties in Wikidata is uncertain.

There might be other constructs in Wikidata that should be considered to be part of the Wikidata ontology. Because of the distributed nature of Wikidata, in general all that can be done to find these constructs is to look through the large amount of ill-structured natural language information about Wikidata. The constructs investigated in this paper are the ones that the authors have encountered as they have investigated the Wikidata ontology and tried to fix issues in it.

This work should also be extended to the rest of Wikidata, in particular to provide an account of temporal qualifiers. This extension will require some sort of temporal model theory and is likely to run into situations where the intended meaning of these qualifiers is uncertain.

Declaration on Generative AI

Aside from viewing the suggestions of AI-guided search such as for determining how best to meet the CEUR guidelines for formatting papers in LaTeX, generative AI tools have not been used in the preparation of this paper. None of the text in this paper was generated by generative AI tools. The technical work described in this paper does fit within AI, notably Knowledge Representation, but not within generative AI.

References

- [1] R. Cyganiak, D. Wood, M. Lanthaler, RDF 1.1 concepts and abstract syntax, W3C Recommendation, www.w3.org/TR/rdf11-concepts, 25 February 2024.
- [2] D. Brickley, R. V. Guha, RDF Schema 1.1, W3C Recommendation, www.w3.org/TR/rdf-schema/, 2014.
- [3] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, P. F. Patel-Schneider (Eds.), The Description Logic Handbook: Theory, implementation, and applications, 2nd ed., Cambridge University Press, 2007.
- [4] B. Motik, P. F. Patel-Schneider, B. C. Grau, OWL 2 web ontology language direct semantics (second edition), W3C Recommendation, www.w3.org/TR/owl2-direct-semantics/, 2012.
- [5] P. Hayes, P. F. Patel-Schneider, RDF 1.1 semantics, W3C Recommendation, www.w3.org/TR/rdf11-mt/, 2014.
- [6] M. Schneider, OWL 2 web ontology language: RDF-based semantics (second edition), W3C Recommendation, <http://www.w3.org/TR/owl-rdf-based-semantics/>, 2012.
- [7] M. Marx, M. Krötzsch, V. Thost, Logic on MARS: Ontologies for generalised property graphs., in: Proceedings IJCAI 2017, 2017, pp. 1188–1194.
- [8] M. Marx, M. Krötzsch, V. Thost, Logic on MARS: Ontologies for generalised property graphs (extended technical report), pdfs.semanticscholar.org/9ba0/36994b7ec73f2c232f08860c76a185755036.pdf, 2017.
- [9] P. F. Patel-Schneider, D. Martin, Wikidata on MARS (preliminary report), in: 33rd International Workshop on Description Logics, 2020. Extended version available as arxiv.org/abs/2008.06599.
- [10] D. Martin, P. F. Patel-Schneider, Wikidata constraints on MARS, in: Wikidata Workshop at The 19th International Semantic Web Conference, 2020. Extended version available as arxiv.org/abs/2008.03900.
- [11] P. F. Patel-Schneider, Reasoning in RDFS is inherently serial, at least in the worst case, in: Proceedings of the Eleventh International Semantic Web Conference, Posters and Demonstrations Track, 2012.