An Analysis of Links in Wikidata

Armin Haller[1 [0000−0003−3425−0780]], Axel Polleres[2 [0000−0001−5670−1146]], Daniil Dobriy[2 [0000−0001−5242−302X]], Nicolas Ferranti[2 [0000−0002−5574−1987]], and Sergio J. Rodríguez Méndez[1 [0000−0001−7203−8399]]

[1] Australian National University, Canberra ACT 2601, AU
{firstname.lastname}@anu.edu.au
[2] Vienna University of Economics and Business, Vienna, AT
{firstname.lastname}@wu.ac.at

Abstract. Wikidata has become one of the most prominent open knowledge graphs (KGs) on the Web. Relying on a community of users with different expertise, this cross-domain KG is directly related to other data sources. This paper investigates how Wikidata is linked to other data sources in the Linked Data ecosystem. To this end, we adapt previous definitions of ontology links and instance links to the terminological part of the Wikidata vocabulary and perform an analysis of the links in Wikidata to external datasets and ontologies from the Linked Data ecosystem. As a side effect, this reveals insights on the ontological expressiveness of meta-properties used in Wikidata. The results of this analysis show that while Wikidata defines a large number of individuals, classes and properties within its own namespace, they are not (yet) extensively linked. We discuss reasons for this and conclude with some suggestions to increase the interconnectedness of Wikidata with other KGs.

1 Introduction

Wikidata, as a "multilingual Wikipedia for data" [25], has grown to a knowledge graph (KG) containing over 95M entities[3]. Since its beginning in 2012, Wikidata has been conceived as a KG that is built bottom-up by its many editors (plus, partially, automatic bots). As a backend, Wikidata uses Wikibase, an open-source software suite for creating collaborative knowledge bases, which allow its many editors to contribute to this KG. Being build bottom-up by domain experts who often also maintain the external original source of data that is being added, Wikidata already includes many links to other datasets, for example, through the reuse of external identifiers for entities (e.g., ORCID records for academics, DOIs for digital artefacts, or the Ensembl identifier for genes (e.g., Q14864292)). This allows the editors of Wikidata to (automatically) integrate data from external KGs that remain under the control of the original publisher. In fact, such automatic integration of external data through bots[4] already exists on Wikidata itself, e.g., a Citationgraph_botthat updates citation numbers of academic works. Consequently, Wikidata has become in practice a *data directory* that serves as entry point to external datasets, other knowledge graphs, or ID providers, respectively. These observations motivate a more in-depth study on the linkage of Wikidata with other KGs and the types of links used for such linking.

[3] cf. https://www.wikidata.org/wiki/Special:Statistics
[4] https://www.wikidata.org/wiki/Wikidata:Bots

Previous work has established link types definitions between datasets [13]. Broadly, this work defined two categories of links, ontology links and instance links. We aim to herein re-use and adapt these definitions and apply them to the Wikidata data model. To do so, we evaluate the HDT dump of the entire Wikidata KG from March 3rd, 2021[5].

For the analysis of ontology links, however, we can not directly use the established link types definitions in [13], since Wikidata does practically not rely on the RDFS/-OWL semantics and vocabularies. While – strictly speaking, in terms of its (RDFS and OWL) TBox constructs used – Wikibase and, as such, Wikidata, use a very simple ontology (i.e., `wikiba.se/ontology`), the *actual* ontology to describe entities in Wikidata is largely build bottom-up by the community itself, not using RDFS/OWL. Indeed, Wikidata partially tries to re-use and integrate external ontologies, but it does so by introducing its own meta-model, and only links to external ontologies through specific, again community-introduced, property relations, such as equivalent class (P1709). This flexibility allows the community to extend the knowledge graph rapidly by adding a rich set of statements about entities in the world without much concern for (logical) consistency expected in the stricter frameworks of RDFS and OWL. This liberty comes with drawbacks, though, with semantic errors or inconsistencies, such as incoherent meta-modeling of classes/instances [23] (i.e., using a taxonomy relation instance of (P31) or a subclass of (P279) relation for similar items[6]), being prevalent. However, many of these problems are eventually resolved through discussions among the editors. There have been some studies on such quality issues within Wikidata [18], but generally there is still little understanding of the quality and evolution of knowledge contained within Wikidata, particularly on the schema level and the schematic relations to other ontologies on the Web.

We therefore present an extension of the definitions of ontology links in [13] by mapping them to the informal, community-developed Wikidata meta-model. In the course of that, we also compare the available meta-properties in Wikidata to their respective corresponding properties in the OWL and RDFS vocabulary which allows us to draw some preliminary conclusions about the ontological expressivity used in Wikidata's meta-modeling. The mapping also allows us to analyse the extend of ontology links and instance links from Wikidata to other KGs. Specifically, we aim to investigate how central Wikidata is to the Linked Data ecosystem by testing the following hypotheses in our analysis.

First, for a KG to serve as a central hub for Linked Data, it should use classes and properties that are defined within its own namespace to represent entities in its KG. Not relying on external ontologies to provide semantics to entities within makes a KG robust to changes in the semantics or availability of external ontologies and as such, a reliable link target for other KGs. It has been observed in our previous study [13] that DBpedia, an existing central hub for Linked Data, exhibits this phenomena that we test in our first hypothesis.

H1 Wikidata defines the vast majority of its terminological entities and properties in its authoritative namespace.

[5] https://www.rdfhdt.org/datasets/

[6] For instance, the pattern { `[] wdt:P279 ?X ; wdt:P31 ?X .` } indicates ambiguous subclass vs. instance of usage on 2131 entities, run on 9 Dec 2021 at `https://w.wiki/4XQw`

Our next set of hypotheses are concerned with the extend to which Wikidata is linked to other ontologies in the Linked Data ecosystem.

H2.1 As a central KG, the ratio of class links to classes defined within Wikidata is much larger than the same ratio for other datasets in the Linked Data ecosystem.

H2.2 As a central KG, the ratio of property links to properties defined within Wikidata is much larger than the same ratio for other datasets in the Linked Data ecosystem.

H3 As a central KG, Wikidata does not type entities using classes from external ontologies, i.e., classes using a namespace other than the authoritative namespace of Wikidata.

Our next two hypotheses are concerned with the extend of which Wikidata is linked to other KGs on an instance level and if the link targets are indeed RDF data.

H4.1 As a central KG, Wikidata includes links from entities defined in its authoritative namespace to entities defined in other KGs and the ratio of such instance links to entities defined in Wikidata is much larger than for other datasets in the Linked Data ecosystem.

H4.2 The amount of instance links to RDF resources is relatively higher than to other types of Web resources, i.e., the content type of the target URI in an instance link is a common RDF serialisation.

In our last hypothesis we test for how many of the entities defined within Wikidata, it is (claims to be) the only authoritative source. A central hub for Linked Data should **not** be the authoritative source for entities, but rather only provide a persistent identity for an entity, while linking to the authoritative external source.

H5 Wikidata establishes equivalence or some weaker forms of likeness relations for the majority of its unique individuals that are part of an instance link, i.e., between entities defined within the Wikidata authoritative namespace and entities defined in other authoritative namespaces.

The remainder of this paper is structured as follows. In Section 2 we discuss the ontology in Wikidata and provide our mapping semantics between the Wikidata meta-model and RDFS/OWL. In Section 3 we describe our methodology to analyse link types in Wikidata. Section 4 presents the results of this analysis and the hypotheses tested on the entire Wikidata RDF corpus. We discuss related work in Section 5 before we conclude in Section 6.

2 The Wikidata Ontology Schema

In terms of a formal backbone terminology, Wikidata relies on Wikibase's minimal predefined schema, i.e., `wikiba.se/ontology` that is used to describe the wiki *pages* of an entity on Wikidata, and, among other things, defines what constitutes a statement for an entity through the `wikiba.se/ontology#Statement` class. However, for our research, this ontology is somewhat irrelevant, as we are looking at internal and external links between entities and the schema (properties and classes used at the statement level) in Wikidata's itself, rather than the Wikibase meta-model. The actual vocabulary used to describe entities in Wikidata is collaboratively built, bottom-up, and indeed its own meta-modelling properties, similar to RDFS/OWL vocabulary properties, have been introduced to this end in the Wikidata namespace. That is, while Wikidata follows the RDF model, it does not use the RDFS or OWL semantics for its ontological

meta-model: it rather conflates[7] what in the traditional Semantic Web stack is defined in RDFS and OWL, i.e., the knowledge about things, groups of things, and relations between things, with what would normally be defined in upper-level or domain ontologies. In this section we will therefore discuss the specific meta-modelling classes and relations that are introduced in Wikidata, their relations and – where possible – their mapping to RDFS/OWL. This mapping will form the basis of our link analysis. We *emphasize* that our proposed mapping is one possible interpretation of the (evolving) meta-model in Wikidata, with the specific purpose of providing formal semantics for our link analysis: we acknowledge that the community does not provide such a mapping by design, in order to avoid (too) strong formal ontological commitment.

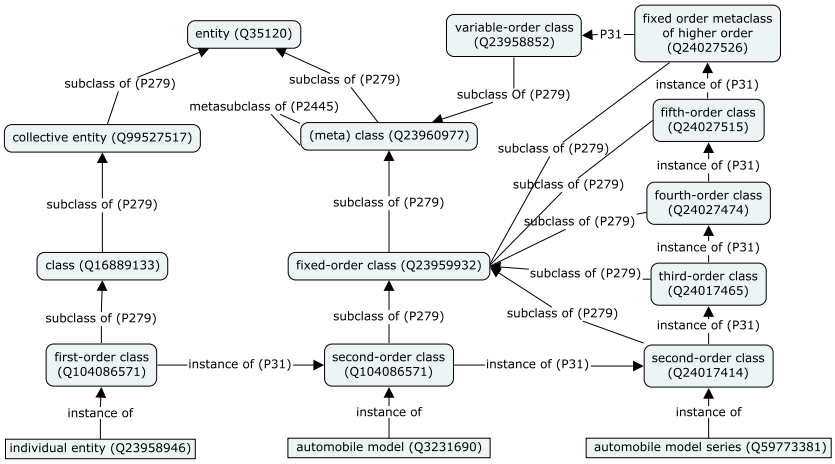

Fig. 1. Overview of the top-level class hierarchy of the Wikidata ontology

2.1 Classes in Wikidata

Figure 1 presents the top-level class hierarchy of Wikidata. Wikidata formally distinguishes between items that are classes, for example, person (Q215627), and items that are instances, for example, Barack Obama (Q76) who is an instance of (P31) human, which itself is a subclass of (P279) person (Q215627). These instances are related to its class via the instance of (P31) relation. Classes in Wikidata are items that are in the object position of an instance of (P31) relation or in the subject or object position in a subclass of (P279) statement. There is also a is meta-subclass of (P2445) relation, but it is hardly used in Wikidata[8]. The Wikidata top-level class for class items is class (Q16889133), which itself is a subclass of entity (Q35120). Wikidata also distinguishes between first-order classes (Q104086571) and second-order classes (Q104086571)[9].

[7] See https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology/Top-level_ontology_list for the top two layers of the ontology

[8] i.e., there are 37 uses of P2445 in total in Wikidata as of August 2021

[9] There are higher orders of second-order class, i.e., third-, fourth- and fifth-order classes, each of which is an instance of the higher ordered class, all of which are subclasses of the fixed-order class (Q23959932).

A metaclass (Q19361238) is defined that is the superclass of fixed-order classes. As such, a second-order class is a metaclass, the instances of which are classes of individuals, for example, the aircraft class (Q1875621) is a second-order class whose members (instances) are first-order classes, including for example wide-body quadjet (Q19394992) and aircraft functional class (Q20027953) which has a subclass wide-body twinjet (Q18683432). While the latter (i.e., the use of an aircraft functional class only for wide-body twinjet, but not for wide-body quadjets) is an example of a non-intuitive subclass hierarchy in Wikidata, it may also reflect different modelling choices by different users. We have argued previously [14] that such a bottom-up development may eventually lead to a more broadly accepted Web ontology. Also, while Wikidata does distinguish between classes and instances, it does not mandate that instances can not also be subclasses of (P279) a class or classes can not be defined as instances of (P31) an instance. While such meta-modeling is not per se forbidden in OWL2 (i.e., through "punning"[10]), in Wikidata it often appears with entities that should be either a class or an instance, but not both. For example, Wiener Schnitzel (Q6497852) is a subclass of schnitzel which is a subclass of meat dish, while at the same time Wiener Schnitzel is also an instance of veal dish which itself is a subclass of meat dish. The only assumption in Wikidata is that entity (Q35120) is the class of all items and therefore all items are an instance of entity (Q35120), as well as all classes are subclasses of entity (Q35120) which is not unlike the role of rdfs:Resource in the RDF meta-model.

2.2 Properties in Wikidata

Properties in Wikidata use the full generality of RDF properties in the sense that they represent both binary (object) relations and (atomic value) attributes. That is, properties are used to define arbitrary item-property-value triples where the value can either be an item or a literal. On top of that, reminding one of RDF's reification mechanism, each such statement can also be qualified, i.e., additional information can be added to the statement (e.g., contextual or provenance information). Indeed, the relation between Wikidata's qualified statements to reification and other potential meta-statement encodings in RDF has been discussed in detail by Hernández et al. [15].

Properties, like entities, have their own Wikidata page and use opaque identifiers starting with "P". Wikidata reuses some RDFS/OWL properties, e.g., rdf:type, rdfs:label, owl:complementOf, owl:someValuesFrom, owl:allValuesFrom etc. However, as discussed earlier, these are merely used to define statements about pages in Wikidata using the wikiba.se ontology, rather than for terminological statements about Wikidata classes and properties. Yet, in order to define such terminological properties within the Wikidata KG, "properties for properties" (73 as of November 2021)[11] are defined in the Wikidata "ontology".

As compared to other RDF KGs, where typically the RDFS/OWL namespaces are used for terminological properties and a separate ontology namespace is used for domain entities/relations, Wikidata's terminological properties use the same namespace as properties that describe entities/relations. For example, in KGs using the RDFS/OWL semantics, relations such as "subClassOf" use the RDFS namespace while a relation like "name" uses a domain-ontology namespace such as FOAF, whereas in Wikidata "subclass of" (P279) and "given name" (P735) share the same namespace.

[10] cf. https://www.w3.org/TR/owl2-new-features/#Simple_metamodeling_capabilities

[11] cf. https://www.wikidata.org/wiki/Wikidata:List_of_properties/Wikidata_property_for_properties

RDFS/OWL property	Equivalence established through	Wikidata property
rdf:type	equivalent property (P1628)	instance of (P31)
rdfs:subClassOf	equivalent property (P1628)	subclass of (P279)
rdfs:subPropertyOf	equivalent property (P1628)	subproperty of (P1647)
rdfs:subPropertyOf	equivalent property (P1628)	external subproperty (P2236)
Inverse rdfs:subPropertyOf	equivalent property (P1628)	external superproperty (P2235)
rdfs:range	equivalent property (P1628)	expressed via property constraint (P2302)
rdfs:domain	equivalent property (P1628)	expressed via property constraint (P2302)
rdfs:label	documented as matching[12]	rdfs:label
rdfs:comment	documented as matching[12]	schema:description
rdf:first	documented as matching[12]	expressed via series ordinal (P1545)
rdf:rest	documented as matching[12]	expressed via series ordinal (P1545)
rdfs:member	documented as matching[12]	part of (P361)
Inverse rdfs:member	inverse property (P1696) of part of (P361)	has part (P527)
owl:equivalentProperty	equivalent property (P1628)	equivalent property (P1628)
owl:equivalentClass	equivalent property (P1628)	equivalent class (P1709)
owl:inverseOf	equivalent property (P1628)	inverse property (P1696)
owl:differentFrom	equivalent property (P1628)	different from (P1889)
owl:unionOf	equivalence intended[13]	union of (P2737)
owl:disjointUnionOf	equivalence intended[13]	disjoint union of (P2738)
owl:onProperty	no documented equivalence	possible candidates: property constraint (P2302)
owl:sameAs	no documented equivalence	possible candidates: exact match (P2888), said to be the same as (P460)
owl:disjointWith	no documented equivalence	N/A
owl:propertyDisjointWith	no documented equivalence	N/A
owl:propertyChainAxiom	no documented equivalence	N/A
owl:assertionProperty	no documented equivalence	N/A

Table 1. Mapping of Wikidata properties to RDFS/OWL properties

In the following we discuss under which circumstances we treat which of these properties as equivalent to their related RDFS/OWL properties as per Table 1: additional to introducing its own already mentioned instance of (P31), and subclass of (P279) relations, to describe taxonomic relations and identify class memberships and hierarchies, respectively, Wikidata introduces a property to describe the hierarchical relation between properties, i.e., subproperty of (P1647), which we consider equivalent to the rdfs:subPropertyOf relation. However, note that this property is used in Wikidata exclusively to link properties within the Wikidata namespace, i.e., it is not used for linking to external vocabularies. For linking to properties external to the Wikidata namespace,

external subproperty (P2236) and external superproperty (P2235) are introduced. We consider these equivalent to rdfs:subPropertyOf, or its inverse, respectively.

Domain and range properties are not directly defined in Wikidata. These property restrictions[14] can be stated in Wikidata using a qualified property constraint (P2302) on the property, where a skolemized IRI is assigned to the entity that is defined as either a type constraint (Q21503250) for the domain classes of the property, or, respectively, as value-type constraint (Q21510865) for the range classes of the property; the respective target class is referenced with the property (P2308). For example, the domain for the date of birth property (P569) is defined to be Human (Q5) (among others) with the following triples[15],

$$t_1 = \texttt{wdt:P569 p:P2302 \_:qe .}$$

$$t_2 = \texttt{\_:qe ps:P2302 wd:Q21503250 .}$$

$$t_3 = \texttt{\_:qe pq:P2308 wd:Q5 .}$$

where _:qe is actually a skolemized blank node with the IRI `wd:statement/P569-F9768BAA-6BB3-4710-A3E1-B6FB9432D372`. Note that for our analysis of links, when only considering whether an external ontology is referenced on a property, we do not need to distinguish if the object that belongs to an external namespace is a domain or range class, i.e., we will not need to check in our SPARQL query below if the target object is a type constraint (Q21503250) or value-type constraint (Q21510865).

In order to state equivalence between two properties, Wikidata introduces the property equivalent property (P1628). Disjointness between properties cannot be stated in Wikidata: in fact, disjointness between properties was proposed by the community (and voted on)[16], but eventually not included. While the reason for its non-inclusion is unclear, it is challenging to maintain disjointness with other properties in a bottom-up created KG where properties can be added arbitrarily.

There is also no relation to define property chain axioms (i.e., owl:propertyChainAxiom), nor is there support for negative property assertions, i.e., a relation similar to owl:assertionProperty does not exist.

As shown above, Wikidata uses the relation property constraint (P2302) to define restrictions on a property. While such a restriction can have more than one triple, it is otherwise very similar to owl:onProperty; as such, for the purpose of our analysis, we consider it equivalent. Restrictions are linked to either a class or property using the OWL properties, owl:onClass, owl:onProperty.

3 Links in Wikidata

Based on the above correspondences of Wikidata's terminological properties with RDF-S/OWL, we are ready to define different link types in the Wikidata data model. Here

[12] cf. https://www.wikidata.org/wiki/Wikidata:Relation_between_properties_in_RDF_and_in_Wikidata

[13] cf. https://www.wikidata.org/wiki/Wikidata:Property_proposal/Archive/48#P2737

[14] We note here again that subtle semantic differences such as constraining (i.e., CWA) vs implicit (i.e., OWA) semantics of certain properties are not relevant for the purpose of our link analysis.

[15] Prefixes are used as follows: wd: <http://www.wikidata.org/entity/>, wdt: <http://www.wikidata.org/prop/direct/>, pq: <http://www.wikidata.org/prop/qualifier/>, p: <http://www.wikidata.org/prop/>, ps: <http://www.wikidata.org/prop/statement/>

[16] https://www.wikidata.org/wiki/Wikidata:Property_proposal/disjoint_with

we will rely on definitions of link types in the RDFS/OWL model defined in our earlier work [13] in terms of resp. SPARQL queries on the Wikidata model[17]. This enables us to directly provide a quantitative and qualitative analysis of the discovered links in terms of the resp. query results.

3.1 Dataset Corpus and authoritative namespaces

In order to analyse links to other datasets in the Wikidata KG, we first need to establish a list of authoritative namespace URIs that are defined by KGs other than Wikidata. For that, we are using the dataset corpus that was defined and published in [13], i.e., the LODCloud, augmented with historically available datasets that were cached in the LODLaundromat [4] and provided as a downloadable corpus in HDT [8]. The Wikidata HDT file (using the http://www.wikidata.org namespace URI) was added to that corpus. The resulting corpus consists of 431 Linked Datasets, each encoded in HDT for a total size of 104 GB (uncompressed 353 GB), with a total number of 17,841,499,814 (i.e.,≈17.8 billion) triples.

3.2 Ontology corpus

As with the dataset corpus, we are reusing the ontology corpus published by Haller et al. [13], i.e., a crawl of the unique classes and properties in `prefix.cc` as well as the declared classes and properties in each dataset (the 431 from above). While not every ontology is registered in prefix.cc (a total of 2,794 ontologies are registered as of August 2021), our process also follows all import statements in those ontologies. Given that ontologies are supposed to be a shared conceptualisation of a domain, if no other ontology reuses the ontology, it is unlikely to be used in many datasets, nor is it likely been used in Wikidata either.

3.3 Link Type Analysis

As per the definitions in our previous work [13] we distinguish two general types of links, Ontology (TBox) Links and Instance (ABox) Links. Ontology Links are further classified into *class links*, *instance typing links*, *property links* and *instance role links*. Instead of providing re-definitions of those links to match the meta-model of Wikidata, we provide reformulations of the operationalised SPARQL queries that retrieve those link types and that implement the mapping relations between the Wikidata meta-model and the RDFS/OWL semantics as defined in Table 1.

For instance, let the dataset ds_{WD} be Wikidata with the set of its authoritative namespaces (i.e., the namespaces denoting Wikidata-defined URIs) being $NS_{ds_{WD}} = \{$ `http://www.wikidata.org/entity/`, `http://www.wikidata.org/prop/direct/`, `https://www.wikidata.org/wiki/Special:EntityData/`$\}$. Further, let ds_2 denote the Disease Ontology, with $NS_{ds_2} = \{$ `http://identifiers.org/doid/`$\}$ and let further ds_3 denote the schema.org vocabulary, with $NS_{ds_3} = \{$`http://schema.org/`$\}$. We shall denote the mentioned namespaces with the prefixes `wd:`, `wdt:`, `data:`, `doid:`, and `schema:`, respectively. If we consider now the triple,

$$t_1 = \texttt{wd:Q84263196 wdt:P2888 doid:0080600 .}$$

[17] All code implemented in Python is available at: https://github.com/arminhaller/LinksInLOD

in ds_{WD}, stating that COVID-19 (Q84263196) is an exact match (P288) to the class DOID:0080600 in the Human Disease Ontology, it shall be considered an *instance link*, from ds_{WD} to ds_2: while doid:0080600 is a class in the Human Disease Ontology, it is used in an instance position in the triple above from Wikidata, therefore it is an *instance link*. The next triple we consider

$$t_2 = \texttt{wd:Q84263196 wdt:P31 wd:Q609748.}$$

defines COVID-19 (Q84263196) as an instance of (P31) of the emerging communicable disease (Q609748) class. While this is not a link, but rather an internal ontological reference within ds_{WD}, however,

$$t_2' = \texttt{wd:Q84263196 rdf:type schema:Dataset.}$$

is indeed an *ontology link*, more specifically, an *instance typing link* from ds_{WD} to ds_3. In fact, every wiki page in Wikidata is defined as of type schema:Dataset. Next,

$$t_3 = \texttt{wdt:P569 wdt:P1628 schema:birthDate .}$$

is an example of a *property link* from ds_{WD} to ds_3. Finally,

$$t_4 = \texttt{wd:Q5 wdt:P1709 schema:Person}$$

is a *class link* from ds_{WD} to ds_3.

In order to define these link types more clearly, in the following we provide SPARQL queries on the Wikidata data model that correspond to the link types defined in [13], adapted to the correspondences in Table 1.

Ontology (TBox) Links With the query shown in Listing 1.1 we retrieve all external classes, i.e., classes using a namespace other than the Wikidata namespace (using the FILTER statement) that are not explicitly declared as an RDFS/OWL class (which no class in Wikidata is) or as a type of class (Q16889133), but are used to *i)* define an instance (i.e., they are used in an assertional axiom), *ii)* define a terminological axiom that either extends or narrows a class through a subclass of relation (P279), *iii)* define a class' equivalence (P1709), union of (P2737), disjoint union of (P2738) or *iv)* define the domain or range of a property (pq:P2308).

Listing 1.1. SPARQL query used to retrieve all external classes.

```
SELECT DISTINCT ?C WHERE {
   {[] a ?C. } UNION
   {[] wdt:P279 ?C. } UNION {?C wdt:P279 []. } UNION
   {?C wdt:P2738 [].} UNION
   {?C wdt:P1709 [].} UNION {[] wdt:P1709 ?C.} UNION
   {?C wdt:P2737 [].} UNION
   {[] pq:P2308 ?C. }
   FILTER (!regex(str(?C), "http://www.wikidata.org","i")) .
}
```

For each class URI retrieved through this query, we check its occurrence in either the subject or object position in any triple in the KG. The number of resulting triples constitutes the number of *Class Links* in the Wikidata KG.

For *Property Links* we follow a similar process. With the query shown in Listing 1.2, we retrieve all external properties (i.e., properties using a namespace other than the authoritative Wikidata namespace) that are not explicitly declared as a property but are used: *i)* within a subproperty relation (P1647) or external sub/superproperty relation (P2236, P2235), *iii)* in a property restriction or to define the domain or range of a class (P2302), or *iv)* to define a properties' equivalence (P1628), inverseness with/to another property (P1696), different from (P1889), complement of (P8882).

Listing 1.2. SPARQL query used to retrieve external properties.

```
SELECT DISTINCT ?P WHERE {
   {?P wdt:P1647 []. } UNION {[] wdt:P1647 ?P. } UNION
   {[] wdt:P2236 ?P. } UNION
   {[] wdt:P2235 ?P. } UNION
   {?P wdt:P1628 []. } UNION {[] wdt:P1628 ?P. } UNION
   {?P wdt:P1696 []. } UNION {[] wdt:P1696 ?P. } UNION
   {?P wdt:P1889 []. } UNION {[] wdt:P1889 ?P. } UNION
   {?P wdt:P8882 []. } UNION {[] wdt:P8882 ?P. } UNION
   {?P wdt:P2302 []. }
   FILTER (!regex(str(?P), "http://www.wikidata.org","i")) .
}
```

For each property URI retrieved through this query, we check its occurrence in the predicate position in any triple in the dataset. The number of resulting property URIs constitutes the occurrence of *Property Links* in the dataset.

Instance Links (ABox Links) Before we can compute the number of *Instance Links* from an individual in the Wikidata namespace to any individual in an external namespace, we first need to find all unique individuals in the KG.

1. We find all individuals of classes/properties that are declared (i.e., individuals that are defined as a type of a class/property using (P31)). For each retrieved unique individual, we check if they are defined in the Wikidata namespace. If not, the triple they appear in is counted as an *Instance Typing Link*.
2. We then find all individuals that are reused from a non-authoritative namespace URI in the subject position without being explicitly declared as a type of a class or property. To retrieve those, we first query all triples in the dataset and then check for each unique subject URI that is not in the Wikidata namespace, if it is already in the set of declared instances (as of step 1), or if it is in the set of classes and properties (cf. Section 3.2). If it is neither, we count the triple as an *Instance Link*.
3. We then follow a similar process for each individual reused from a non-authoritative namespace URI in the object position. For each unique object URI, we check the following conditions: *i)* the subject URI does not contain the Wikidata namespace URI, *ii)* the predicate is not an instance of (P31) relation, and *iii)* the object URI is not already contained within the set of declared instances. If none of these conditions are satisfied, we record it as an *Instance Link*.

4 Evaluation of Links

In the following, we discuss the results of the link analysis of Wikidata. All tests were performed on a machine with 8vCPUs, 380GB RAM and 5TB hard disk space.

4.1 General statistics of the Wikidata KG

Before we analyse the number and types of links in Wikidata, we present some general statistics of the Wikidata KG that we computed using its HDT file in Table 2. The first noteworthy observation we can make is that the ratio between unique subjects to unique predicates in Wikidata is 1/41810, whereas in the LOD dataset corpus [13] the ratio was 1/3900 if using the mean and 1/19 if using the median. Since the Wikidata KG with 1.69bn triples is much larger than the largest KG in the LOD corpus, i.e., the 2016 version of DBpedia with 1.04bn triples, which itself was much larger than the mean number of triples (i.e. 16.92m) for all datasets in the corpus, suggests that the number of predicates in a KG grows following an asymptotic function. This seems natural, as while the number of entities in a general KG such as Wikidata is potentially infinite, the attributes that can be assigned to those entities are somehow limited.

Supporting our first hypothesis, Wikidata defines all of its 89m unique individuals using its own ontology. The number of unique individuals is also larger than the number of claimed unique entities defined in Wikidata at the time the HDT file was generated (March 2021), i.e., 73m, meaning that all entities (plus some more) are defined within the Wikidata namespace. While the number of unique subjects, with 1.62bn is a lot larger than the number of unique individuals (i.e., 89m), this is due to the fact that Wikidata introduces skolemized IRIs for qualified statements on an entity, and not because of entities being defined using external ontologies, i.e., instance typing links (see below). As such, on average for each unique entity about 18 entities are created as part of the subgraph for that entity and that redirect in the Linked Data API to that target entity. For example, for the entity Barack Obama (Q76) there are 394 skolemized IRIs (as of November 2021), such as http://www.wikidata.org/entity/statement/q76-F23589FF-58A6-438B-BC7E-79F6B436AFD0 that describes a qualifier about Barack Obama's education at (P69) Harvard Law School which he completed with an academic degree (P512) of Juris Doctor with an end time (P582) of 1991. These qualified statements do not have their own page on Wikidata, but they do resolve to the page where they are defined through the Linked Data API.

# Triples	1,693,668,039
# Unique Subjects	1,625,057,179
# Unique Predicates	38,867
# Unique Objects	2,538,585,808

Table 2. General statistics of the Wikidata KG

# Unique Individuals	89,120,227
# Unique Declared Classes	0
# Unique Undeclared Classes:	2,522,595
# Unique Declared Properties:	74,309
# Unique Undeclared Properties:	29,167

Table 3. Class/property statistics in Wikidata

4.2 Ontology Links

Before we set out to test our hypotheses related to ontology links in Wikidata, we first present some general statistics on the use of classes and properties in Wikidata in Table 3. For our analysis we distinguish between declared and undeclared classes, i.e., class URIs that are defined within the authoritative namespace of the KG using a triple {`[] rdf:type owl:Class .`} or {`[] rdf:type rdfs:Class .`}, and class URIs that are merely reused from a different namespace URI. Since Wikidata does not use rdf:type relations for class definitions (as above), all 2,522,595 unique classes defined in Wikidata are undeclared according to the RDFS/OWL semantics.

In contrast to class URIs, properties in Wikidata are declared using the owl:DatatypeProperty and owl:ObjectProperty types. In fact, each property, denoted by an identifier starting with "P" includes up to nine datatype and object property definitions, each with a different URI of that property identifier (i.e., strictly speaking different properties) as defined by the wikiba.se ontology, e.g., for date of birth (P569) this includes http://www.wikidata.org/prop/direct/P569 which is defined as an owl:DatatypeProperty and http://www.wikidata.org/prop/statement/P569 as a shorthand property to find the statements this property is used in which is also defined as an owl:DatatypeProperty. Therefore the number of declared properties in March 2021 (i.e., 74,309) is more than six times larger than the claimed number of properties on Wikidata, i.e., 9,367 properties as of November 2021[18][19].

# Class Links	3,955
# Property Links	835
# Instance Typing Links	0 (173,168,537)
# Instance Links	173,177,045

Table 4. Link Type statistics

Class Links There are only 3,955 class links defined in Wikidata. This is comparable to the other datasets in the LOV corpus that were analysed previously [13], where the number of class links is relatively constant around 100-10,000 per dataset. However, Wikidata uses 2.5m classes (compared to an average of 6,379 classes per dataset), with a ratio of class links per class of only 0.0015, while the average ratio of class links per class for the datasets in the LOV cloud is 11.27. One of the reasons why the ratio is so low, is that many instances in Wikidata are also defined as classes (see above), contrary to many other KGs where there is a strict separation between TBox and ABox axioms. Also, the user interface's (Wikibase) autocomplete feature when creating links to classes only works for classes in the Wikidata namespace, but not for external URIs of classes. Still, as a central hub of the Linked Data ecosystem, one would expect Wikidata to have more such links, particularly given the bottom-up development of the Wikidata ontology. The 2016 version of DBpedia [2], for example, includes 8,258 class links for its 3,197 classes with a ratio of 2.58, even though its ontology is built by experts top-down. We therefore need to reject our hypothesis H2.1. To increase the number of class links, a relation similar to "external subproperty" should be introduced in Wikidata to define external subclass relations on a class. A lookup service (based for example on the LOV API [24]) could then guide Wikidata editors to the existence of external classes.

Property Links There are only 835 property links for a total of 74,309 properties in Wikidata, i.e., there are on average only 0.01 property links per property, a ratio that is much lower than for the LOD corpus. We therefore must also reject our hypothesis H2.2, that Wikidata includes many more property links per property than other datasets. One of the reasons for this low property link ratio might be that while there exist several properties in the Wikidata ontology that are specifically designed to link to external ontologies or allow external URIs to be used, i.e., equivalent property (P1628), different

[18] https://www.wikidata.org/wiki/Wikidata:List_of_properties

[19] No longitudional data is published on the Wikidata site, but the growth in the number of properties between July and November 2021 was 3.4%

from (P1889), external subproperty (P2236), external superproperty (P2235) these are only relatively recent additions. External subproperty and superproperty which are used 94 and 159 times, respectively, were only added in May 2017 and May 2018, after many of the 74k properties in Wikidata have been defined. There would need to be a concerted effort by the community to update existing properties with these relations.

Instance Typing Links There are no instance typing links in Wikidata that use the instance of (P31) relation. Since, to the best of our knowledge, Wikibase does not allow users to add an external URI when using the instance of relation it is not unexpected that there are no such links. There are 173m instance typing links using the rdf:type relation. However, since they are all (auto-generated) links to define a Wikidata page as a schema:Dataset and a schema:Article we excluded them. Therefore we can confirm our hypothesis H3 that as a KG with a general and sufficiently comprehensive ontology, Wikidata types all entities using its own ontology and therefore includes no instance typing links.

4.3 Instance Links

Wikidata includes many links from entities (unique individuals) defined in its authoritative namespace to entities defined in other KGs. With 173m such links, it means that 10.22% of all triples in the Wikidata graph link to individuals that use a namespace other than the Wikidata namespace. However, the ratio of such links to entities (at 1.94) is much lower than with other datasets in the LOD ecosystem (8.6) and we therefore need to reject our hypothesis H4.1.

Even this number includes many links to Wikipedia. In fact, every entity in Wikidata that also has a Wikipedia entry includes hundreds of links to Wikipedia. However, while they are considered links according to our definition, none of the target resources are, in fact, RDF resources, but the Wikipedia entity is created in the Wikidata namespace (using the Wikipedia URL).

While for ontology links we are able to verify for all links if the target URI is an RDF resource, with the large number of instance links, we can not. However, to test our hypothesis H4.2 a sample of 1,924,940 target URIs from all instance links was randomly collected. We then built a simple crawler that checked for each URI if a document in RDF format can be retrieved at the target URI, i.e., classifying the links in two main groups: Web resources, and RDF entities. Table 5 shows that the majority of links point to resources other than RDF. While we therefore need to reject our hypothesis H4.2, the fact that a quarter of resources are, in fact, RDF resources is encouraging, given that for many entities there may not yet exist an RDF representation outside of Wikidata. To distinguish RDF from non-RDF resources in links, i.e., to distinguish 1-star linked data from higher-ordered linked data [5], Wikidata should automatically qualify links based on the target format of the linked resource.

# URL Not Found	18,081 (0.9%)
# Other Errors	138,656 (7.2%)
# Timeout	218,542 (11.4%)
# RDF Entities	471,088 (24.5%)
# Web Resources	1,078,573 (56.0%)

Table 5. Instance links content-types statistics

# owl:sameAs Links	0
# Exact Match (P2888) Links	3,268,021
# Said to be the Same (P460) Links	2
# Inverse Property (P1696) Links	0

Table 6. Instance Link types statistics

We also checked how many of the instance links use an equivalence or some weaker forms of likeness relations to test our Hypothesis 5. Unsurprisingly, no instance link uses the owl:sameAs relation, as Wikibase does not allow its use and encourages the use of the exact match (P2888) relation. However, with 3,268,021 such links, at most [20] only 3.7% of all unique individuals use the exact match relation to an individual defined in a namespace other than the Wikidata namespace. P460 and P1696 are not used. We therefore must also reject Hypothesis 5.

H1	Wikidata defines the vast majority of its terminological entities and properties in its authoritative namespace	**Supported**
H2.1	The ratio of class links to classes in Wikidata is higher than in the LOD ecosystem	**Rejected**
H2.2	The ratio of property links to properties in Wikidata is higher than in the LOD ecosystem	**Rejected**
H3	Wikidata does not type entities using classes from external ontologies	**Supported**
H4.1	Wikidata's ratio of instance links to entities is higher than for other datasets in the LOD ecosystem	**Rejected**
H4.2	Most instance links point to RDF Web resources	**Rejected**
H5	Wikidata includes similarity relations for a majority of its instance links	**Rejected**

Table 7. Hypotheses testing

5 Related Work

There are many works that analyse different quality aspects of Wikidata. Erxleben et al. [9] introduce RDF exports that connect Wikidata to the Linked Data Web. In [6] an axiomatic theory for multi-level modeling is used to analyse Wikidata content and to identify a significant number of problematic classification and taxonomic statements. Färber et al. [11] present an extensive survey of open KGs, including Wikidata. Freire & Isaac [10] present an assessment of Wikidata for high-quality machine interpretation of its alignment properties to RDF/S, OWL, SKOS, and schema.org.

Piscopo & Simperl [18] present a systematic literature review of 28 papers about data quality in Wikidata, categorised by quality dimensions addressed. The completeness aspect of Wikidata is analysed in [3], which cites some tools and services that address various quality aspects around the WikiMedia projects. Pillai et al. [16] compare Wikidata with other KGs from the perspectives of completeness of its relations, timeliness of the data, and accessibility as the data quality criteria. Abian et al. [1] present an approach based on cross-comparing date values (the concept of contemporary constraint) to discover inconsistent temporal data in Wikidata. Piscopo & Simperl [17] study the relationship between different Wikidata user roles and the quality of the Wikidata ontology by proposing a framework to evaluate the ontology as it evolves. Samuel [21] introduces the `WDProp` tool that provides to human users an overview and statistics of various multi-language aspects of Wikidata properties, such as labels, descriptions, and aliases. Shenoy et al. [23] present a quality analysis of Wikidata focusing on correctness.

Other work exists that analyse interlinking in linked data in general [26,20] or quality studies and approaches that considered interlinking of linked data as an assessment metric [12,19,7,22].

[20] some individuals might use more than one exact match relation

None of the above works, however, have analysed how interlinked and central Wikidata is to the LOD ecosystem, and more specifically, analysed the number and types of links defined within Wikidata as presented in this paper.

6 Conclusion

We have analysed the number and types of links in Wikidata to evaluate how central Wikidata is to the Linked Open Data ecosystem. While Wikidata is the largest, most comprehensive general knowledge KG on the Web using also a comprehensive, bottom-up developed ontology that is used to type its many entities, it is not (yet) serving as a central hub for linked data on the Web.

For its relative lack of instance links, this means that either the Wikidata editors deem such links as obsolete, or that these links are yet to be included or that they already exist, but rather as incoming links from the external dataset to Wikidata. However, as a bottom-up created KG, there is the possibility for anyone who owns a dataset to actually create an outgoing link in the Wikidata namespace to the dataset they own. Many bots (332, https://www.wikidata.org/wiki/Wikidata:Bots) have been created for exactly this reason (i.e., automatically creating outgoing links from Wikidata to other datasets), and they improve the discoverability, and as such the visibility, of the external dataset. Every dataset publisher, for their own benefit, should therefore consider creating those outgoing links in the Wikidata KG.

Comparatively for its size, Wikidata also includes less ontology links than other datasets in the LOD ecosystem. While this can be partially explained by the fact that many individuals defined in Wikidata are also classes, skewing the ratio between classes and class links, this does not apply to property links, where there is no such distinction. Most properties in Wikidata are not linked to external properties at all, even though specific properties exist in the Wikidata ontology (e.g., external subproperty, external superproperty) to do so. While we have suggested in this paper that some changes to the user interface of Wikibase may encourage editors to provide more such links, a fundamental rethink of ontology design may have to occur too. Specifically, common best-practise in ontology engineering is to include links from an ontology to other ontologies (i.e., through import statements or URI reuse). However, in the case of the Wikidata ontology, the developers of domain ontologies should consider to create those links to their ontologies in the Wikidata namespace, rather than in the other direction.

Wikidata also does not (yet) provide many equivalence or weaker forms of likeness relations from its entities to external entities. There is an onus on the Wikidata editor community to ensure that such links are increasingly provided, given that Wikidata should generally not be the authoritative source of entities, but link to an authoritative representation of an entity through, for example, the exact match relation in Wikidata. However, as above, the lack of such links may also be an indication that entities defined in Wikidata do not yet exist or never will exist in the LOD ecosystem.

As future works, we first would like to analyse the evolution of links on Wikidata over time using several historical snapshots of the published Wikidata HDT files. Also, a deeper analysis of the entities that are linked (e.g., what are the top-ranked instance and ontology namespaces referenced from Wikidata) is planned for a future work.

Acknowledgment: This research has received funding from the Teaming.AI project, which is part of the European Union's Horizon 2020 research and innovation program under grant agreement No 957402.

References

1. Abián, D., Bernad, J., Trillo, R.: Using Contemporary Constraints to Ensure Data Consistency. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. pp. 2303–2310 (April 2019)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Proceedings of the International Semantic Web Conference (ISWC), pp. 722–735. LNCS, Busan, South Korea (2007)
3. Balaraman, V.; Razniewski, S.; Nutt, W.: Recoin: Relative Completeness in Wikidata. In: Wiki Workshop 2018 co-located with The Web Conference 2018 in Lyon, France, April 24, 2018) (April 2018)
4. Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: LOD laundromat: a uniform way of publishing other people's dirty data. In: Proceedings of the International Semantic Web Conference (ISWC). pp. 213–228. LNCS, Riva del Garda, Italy (2014)
5. Berners-Lee, T.: Linked Data. W3C Design Issues (July 2006), from `http://www.w3.org/DesignIssues/LinkedData.html`
6. Brasileiro, F., Almeida, J.a.P.A., Carvalho, V.A., Guizzardi, G.: Applying a Multi-Level Modeling Theory to Assess Taxonomic Hierarchies in Wikidata. In: Proceedings of the 25th International Conference Companion Volume on World Wide Web. p. 975–980 (2016)
7. Debattista, J., Auer, S., Lange, C.: Luzzu - A Methodology and Framework for Linked Data Quality Assessment. Journal of Data and Information Quality **8**(1), 4:1–4:32 (Oct 2016)
8. Debattista, J., Lange, C., Auer, S., Cortis, D.: Evaluating the Quality of the LOD Cloud: An Empirical Investigation. Semantic Web **9**(6), 859–901 (2018)
9. Erxleben, F., Günther, M., Krötzsch, M., Méndez, J., Vrandečić, D.: Introducing Wikidata to the Linked Data Web. In: Proceedings of the International Semantic Web Conference (ISWC) 2014. pp. 50–65. Springer, Riva del Garda, Italy (2014)
10. Freire, N., Isaac, A.: Technical Usability of Wikidata's Linked Data: Evaluation of Machine Interoperability and Data Interpretability. In: 22nd International Conference on Business Information Systems (BIS). Seville, Spain (2019)
11. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semantic Web **9**(1), 77–129 (2018)
12. Guéret, C., Groth, P., Stadler, C., Lehmann, J.: Assessing Linked Data Mappings Using Network Measures. In: Proceedings of the 9th Extended Semantic Web Conference (ESWC). LNCS, vol. 7295, pp. 87–102. Springer, Heraklion, Greece (2012)
13. Haller, A., Fernández, J.D., Kamdar, M.R., Polleres, A.: What are Links in Linked Open Data? A Characterization and Evaluation of Links between Knowledge Graphs on the Web. Journal of Data and Information Quality **12**(1) (2020)
14. Haller, A., Polleres, A.: Are we better off with just one ontology on the Web? Semantic Web **11**(1) (2020)
15. Hernández, D., Hogan, A., Krötzsch, M.: Reifying RDF: what works well with wikidata? In: Proceedings of the 11th International Workshop on Scalable Semantic Web Knowledge Base Systems. vol. 1457, pp. 32–47. CEUR-WS.org (2015)
16. Pillai, S.G., Soon, L.K., Haw, S.C.: Comparing DBpedia, Wikidata, and YAGO for Web Information Retrieval. In: Piuri, V., Balas, V.E., Borah, S., Syed Ahmad, S.S. (eds.) Intelligent and Interactive Computing. pp. 525–535. Singapore (2019)
17. Piscopo, A., Simperl, E.: Who Models the World?: Collaborative Ontology Creation and User Roles in Wikidata. Proceedings of ACM Human-Computer Interaction **2**(CSCW), 141:1–141:18 (2018)
18. Piscopo, A., Simperl, E.: What We Talk about When We Talk about Wikidata Quality: A Literature Survey. In: Proceedings of the 15th International Symposium on Open Collaboration. New York, NY, USA (2019)

19. Raad, J., Beek, W., van Harmelen, F., Pernelle, N., Saïs, F.: Detecting erroneous identity links on the web using network metrics. In: Proceedings of the International Semantic Web Conference (ISWC). pp. 391–407. Springer International Publishing, Cham (2018)
20. Radulovic, F., Mihindukulasooriya, N., García-Castro, R., Gómez-Pérez, A.: A comprehensive quality model for Linked Data. Semantic Web **9**(1), 3–24 (2018)
21. Samuel, J.: Towards Understanding and Improving Multilingual Collaborative Ontology Development in Wikidata. In: Proceedings of Wiki Workshop 2018 co-located with The Web Conference 2018. Lyon, France (April 2018)
22. Sarasua, C., Staab, S., Thimm, M.: Methods for intrinsic evaluation of links in the web of data. In: Proceedings of the International Semantic Web Conference. pp. 68–84. Springer International Publishing, Cham (2017)
23. Shenoy, K., Ilievski, F., Garijo, D., Schwabe, D., Szekely, P.: A study of the quality of wikidata. arXiv preprint arXiv:2107.00156 (2021)
24. Vandenbussche, P., Atemezing, G., Poveda-Villalón, M., Vatant, B.: Linked open vocabularies (LOV): A gateway to reusable semantic vocabularies on the web. Semantic Web **8**(3), 437–452 (2017)
25. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communication of the ACM **57**(10), 78–85 (2014)
26. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for Linked Data: A Survey. Semantic Web **7**(1), 63–93 (2016)