Understanding Wikidata

Andrea Westerinen, 16 March 2024, V1.0

Overview

Wikidata is the largest, open, general-purpose and multi-lingual knowledge base currently available. It holds information on over 100M entities[1] (both concepts, lexemes and instances), and is accessible and editable by both humans and computers. This makes Wikidata extremely valuable as a data source to supplement, validate and extend existing knowledge bases and applications. To effectively utilize Wikidata in a Semantic Web application, one must understand its design, terminology and correspondence with ontological concepts (especially RDF and OWL). Explaining these items is the purpose of this paper.

Each of the following sections discuss:

- Wikidata's data model and terminology
- Mapping of Wikidata entities to RDF/OWL
- Wikidata's ontologies and "frequently used" concepts

1 https://www.wikidata.org/wiki/Wikidata:Data_access/en#:~:text=Wikidata%20currently%20contains%20over%20100,method%20to%20suit%20their%20needs

- Constraints and quality control

A future document will address the correspondence of Wikidata to other, external ontologies and use of the information below to create and reason with an RDF/OWL encoding for Wikidata.

Table of Contents

Understanding Wikidata 1

Overview 1

Table of Contents 2

Wikidata Data Model and Terminology 3

Use of OWL in the Wikidata RDF 13

RDF Triples in Wikidata Linked Data, Dump and WDQS 15

Types of Wikidata Property Values 17

Mapping of Wikidata to RDF/OWL 22

Wikidata Ontologies 24

Frequently Used Wikidata Classes and Properties 29

Alignment with External Ontologies 32

Wikidata Constraints and Quality Control 35

Wikidata Data Model and Terminology

Wikidata's data model[2] resembles the Resource Description Framework (RDF) syntax in that specific pieces of information (identified by "properties") are accessible for an "item". In fact, there is sufficient correspondence to support a Wikidata RDF output and SPARQL Query endpoint. However, it is important to understand that there are differences between standard usage of RDF/OWL in industry and the Wikidata data model and its RDF encoding. This is due to the fact that Wikidata is natively stored in a document-oriented database based on MediaWiki. Its RDF output is only one possible mapping.

When considering Wikidata in an RDF/OWL context, it is important to remember that many (but not all) Wikidata "items" are specifically defined to be concepts or instances that are described in Wikipedia articles. As such, Wikidata's knowledge and definition may be overly specific, repetitive, contradictory or incomplete when examined through an ontological lens.

In its simplest form, the Wikidata data model reflects a basic RDF subject–predicate–object triple. Wikidata items and properties are identified using a naming format consisting of an alphabetic character ("Q" for items, and "P" for properties) followed by an integer. (And, anything with a "Q" or "P" identifier may also be referenced as a Wikidata "entity".)

The actual underlying data is more complex. Wikidata's information about an item is contained in "statements" which involve *not only* properties and their values, but also "qualifiers". The latter provide clarifying information, such as the start or end time of an event, how the object value or reference was determined, additional descriptive detail, and more. Also, supporting references are encouraged to be specified for the statement, as well as its "rank" (e.g., preferred, normal or deprecated).

Taking all the statements into account, the value of the property with the highest "rank" will be returned if a simple subject-predicate-object response is requested. This is accomplished by requesting the "truthy" value of the property (identified in a SPARQL query using the

[2] https://www.mediawiki.org/w/index.php?title=Wikibase/DataModel, https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer

namespace prefix, "wdt:"; more on that in a few paragraphs). Wikidata statements can be viewed as analogous to RDF reification.

The following table summarizes the correspondence between the Wikidata data model and RDF.

Wikidata Terminology	RDF Terminology
Item	Resource (both instances and classes)
Property	Property (both datatype and reference properties)
Qualifier	Property + value pair (analogous to RDF* edge properties)
Statement	Identifier that is the subject of a set of triples defining property-value pairs, with qualifiers and supporting references

Table 1. Correspondence between Wikidata and RDF Modeling Concepts

Veering further from the RDF standard, Wikidata considers "namespaces" (URIs which define the different types of pages in Wikidata – such as help pages, https://www.wikidata.org/wiki/Help:Namespaces) distinct from "SPARQL prefixes" (which are used in SPARQL queries issued to the Wikidata Query Service (WDQS) or another hosting service). This may be confusing to RDF/OWL practitioners since one would expect to consistently use the same "namespace/prefix" for a property. This is *not* a valid assumption in Wikidata.

The figure below shows an example of the Wikidata RDF entity output (the "Linked Data Interface") for an instance (Christopher A. Armatas, Q100049183) of the class, human Q5. It illustrates the concepts of Item, Property, Qualifier and Statement (from the table above) and the use of SPARQL prefixes. The figure is a subset of the RDF/OWL output for the URL, https://www.wikidata.org/entity/Q100049183.ttl. (Note that the "Linked Data" output differs

from the Wikidata RDF Dump output for the complete data because it uses a few different namespace prefixes. For more information, see the first bullet in the subsection, RDF Triples in Wikidata Linked Data, Dump and WDQS.)

```
wd:Q5 a wikibase:Item ;   # "Human" class
   rdfs:label "human"@en ; skos:prefLabel "human"@en ; schema:name "human"@en ;
   schema:description "any member of Homo sapiens, ..."@en .

wd:Q100049183 a wikibase:Item ;   # Instance of person, Christopher Armatas
   wdt:P31 wd:Q5 ;                # "Truthy" triple indicating that the item is an instance of (P31) human (Q5)
   p:P31 s:Q100049183-fe12353d-4d03-aa99-ff8d-6eaceda13cfa ;   # Reference to the full statement for P31
   wdt:P21 wd:Q6581097 ;          # "Truthy" triple indicating the person's gender (P21), male (Q6581097)
   p:P21 s:Q100049183-30993b10-45dc-61f7-4d9a-655db209c0dd ;   # Reference to the full statement for P21
   ...
   rdfs:label "Christopher A. Armatas"@en ; skos:prefLabel "Christopher A. Armatas"@en ;
   schema:name "Christopher A. Armatas"@en ;
   ...
   schema:description "onderzoeker"@nl, "Forscher"@de, "研究員"@ja, "researcher"@en, "chercheur"@fr, ... ;
   skos:altLabel "Christopher Armatas"@en .

s:Q100049183-fe12353d-4d03-aa99-ff8d-6eaceda13cfa a wikibase:Statement, wikibase:BestRank ;
   wikibase:rank wikibase:NormalRank ;
   ps:P31 wd:Q5 .                 # Statement property for instance of (P31) human (Q5)

s:Q100049183-30993b10-45dc-61f7-4d9a-655db209c0dd a wikibase:Statement, wikibase:BestRank ;
   wikibase:rank wikibase:NormalRank ;
   ps:P21 wd:Q6581097 .           # Statement property for gender (P21) male human (Q6581097)

wd:Q6581097 a wikibase:Item ;
   rdfs:label "male"@en ; skos:prefLabel "male"@en ; schema:name "male"@en ;
   schema:description "to be used in \"sex or gender\" (P21) to indicate ... a male person"@en .
```

Figure 1. Example RDF output for the Items Q5 and Q100049183

The remainder of this section further explains the various SPARQL prefixes and their usage.

As mentioned above, related to "truthy" property values, Wikidata distinguishes the use of properties with items as their subjects, versus use of those same properties with statements as their subjects, based on the SPARQL prefix of the property. In RDF/OWL, a property URI would be specified as a namespace and local name, and its definition would include the specification of the domains (subject classes) and ranges (value datatypes or classes) to which the property applies. However, in Wikidata, a property's domain (subject) applicability is specified by its SPARQL prefix, and only its range (value) is included in its definition. And, in Wikidata, SPARQL prefixes also define other distinctions.

A list of SPARQL prefixes is found in the "Prefixes Used" section of the RDF Dump Format web page[3]. In that section, usage is explained, as well as in a succeeding section, Predicates. The following table combines and corrects information from both, and adds details about other prefixes (such as the Blazegraph-specific prefixes associated with the Wikidata query engine).

Prefix	Full URI	Semantics and Usage
bd:	`http://www.bigdata.com/rdf#`	Prefix for the Blazegraph-specific slice and sample functions[4], as well as for specifying SPARQL function parameters (e.g., bd:serviceParam)
cc:	`http://creativecommons.org/ns#`	Prefix of the Creative Commons namespace[5] used with the property, license, to indicate the type of license of Linked Data or RDF Dump output
data:	`http://www.wikidata.org/wiki/Special:EntityData/`	Prefix for a subject node in Linked Data's RDF, whose triples define the entity's metadata (e.g., triples for Douglas Adams Q42 are specified with the subject node, data:Q42). Note that the RDF Dump prefix is wdata:.
dct:	`http://purl.org/dc/terms/`	Prefix for Dublin Core metadata namespace[6] used with the property, language, to indicate the language of a lexeme
gas:	`http://www.bigdata.com/rdf/gas#`	Prefix for the Blazegraph-specific "Gather, Apply, and Scatter" (GAS)[7] service

[3] However, the linked page does not correctly report Linked Data prefixes, instead indicating that they are the same as the RDF Dump output.
[4] https://wikitech.wikimedia.org/wiki/User:AndreaWest/Blazegraph_Features_and_Capabilities#bd:slice and https://wikitech.wikimedia.org/wiki/User:AndreaWest/Blazegraph_Features_and_Capabilities#wikibase:mwapi,_gas:service_and_bd:sample
[5] https://creativecommons.org/ns#
[6] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[7] https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/query_optimization#GAS_Service

Prefix	Full URI	Semantics and Usage
geo:	`http://www.opengis.net/ont/geosparql#`	Prefix of the GeoSPARQL namespace[8] used for properties having a "GlobeCoordinate" data value (such as coordinate location, P625). The value is written as a WKT-compliant string (e.g., "Point(35.3, 12.93)"^^geo:wktLiteral).
hint:	`http://www.bigdata.com/queryHints#`	Prefix to define Blazegraph-specific query hints[9]
ontolex:	`http://www.w3.org/ns/lemon/ontolex#`	Prefix of the Lexicon Model for Ontologies (LEMON) namespace[10] with which the Wikidata Lexeme Model is aligned (see the discussion on the Wikibase:Lexeme/Data Model web page)
owl:	`http://www.w3.org/2002/07/owl#`	Prefix for the W3C OWL standard[11] which is used for several types and predicates (please see a short discussion on the use of OWL at the end of this section)
p:	`http://www.wikidata.org/prop/`	Prefix for a property relating a Wikidata item (a concept or instance, the domain, a node specified with the prefix, wd:) to a statement node (the range, a node specified with the prefix, s: or wds:), where the statement's triples provide information about the property's value and its qualifiers, rank and references

[8] http://www.opengis.net/ont/geosparql#
[9] https://github.com/blazegraph/database/wiki/QueryHints
[10] https://www.w3.org/2016/05/ontolex/
[11] https://www.w3.org/TR/owl2-overview/

Prefix	Full URI	Semantics and Usage
pq:	`http://www.wikidata.org/prop/qualifier/`	Prefix for a property relating an item's statement (the domain, a node specified with the prefix, s: or wds:) to a qualifier value (the range), where the value is a literal or IRI. For example, for a statement about a person's occupation P106, a possible qualifier is start time P580. pq:P580 is then used to specify the explicit start date of the occupation.
pqn:	`http://www.wikidata.org/prop/qualifier/value-normalized/`	Prefix for a property relating an item's statement (the domain, a node specified with the prefix, s: or wds:) to a "normalized" qualifier value (the range), where the qualifier value is a numeric datatype defined using standard SI units[12] or is a reference specified as a full URI
pqv:	`http://www.wikidata.org/prop/qualifier/value/`	Prefix for a property relating an item's statement (the domain, a node specified with the prefix, s: or wds:) to a qualifier value node (the range, a node specified with the prefix, v: or wdv:). The value node is an instance of type, wikibase:TimeValue, wikibase:QuantityValue, etc., with its own set of qualifiers.
pr:	`http://www.wikidata.org/prop/reference/`	Prefix for a property relating a reference node (the domain, a node specified with the prefix, ref: or wdref:) to a reference value (the range), where the value is a literal or IRI

[12] https://en.wikipedia.org/wiki/International_System_of_Units

Prefix	Full URI	Semantics and Usage
prn:	`http://www.wikidata.org/prop/reference/value-normalized/`	Prefix for a property relating a reference node (the domain, a node specified with the prefix, ref: or wdref:) to a "normalized" reference value (the range), where the value is defined using SI units or as a full IRI
prov:	`http://www.w3.org/ns/prov#`	Prefix of the W3C Provenance ontology[13] used with the property, wasDerivedFrom, to indicate the source of a statement and its property-value pair
prv:	`http://www.wikidata.org/prop/reference/value/`	Prefix for a property relating a reference (the domain, a node specified with the prefix, ref: or wdref:) to a value node (the range, a node specified with the prefix, v: or wdv:). The value node is an instance of type, wikibase:TimeValue, wikibase:QuantityValue, etc., with its own set of qualifiers.
ps:	`http://www.wikidata.org/prop/statement/`	Prefix for a property relating an item's statement (the domain, a node specified with the prefix, s: or wds:) to the property's value (the range), where the value is a literal or IRI
psn:	`http://www.wikidata.org/prop/statement/value-normalized/`	Prefix for a property relating an item's statement (the domain, a node specified with the prefix, s: or wds:) to a "normalized" property value (the range), where the value is defined using SI units or as a full IRI

[13] https://www.w3.org/TR/prov-o/

Prefix	Full URI	Semantics and Usage
psv:	`http://www.wikidata.org/prop/statement/value/`	Prefix for a property relating an item's statement (the domain, a node specified with the prefix, s: or wds:) to a value node (the range, a node specified with the prefix, v: or wdv:). The value node is an instance of type, wikibase:TimeValue, wikibase:QuantityValue, etc., with its own set of qualifiers.
ref:	`http://www.wikidata.org/reference/`	Prefix for a provenance reference node in Linked Data's RDF. Note that the RDF Dump prefix is wdref:.
rdf:	`http://www.w3.org/1999/02/22-rdf-syntax-ns#`	Prefix for the W3C RDF standard[14] which is used with the property, type (also written using syntactic shorthand as "a")
rdfs:	`http://www.w3.org/2000/01/rdf-schema#`	Prefix for the W3C RDFS standard[15] which is used with the property, label
s:	`http://www.wikidata.org/entity/statement`	Prefix for a statement node in Linked Data's RDF. Note that the RDF Dump prefix is wds:.
schema:	`http://schema.org/`	Prefix of the schema.org namespace[16] that provides details for RDF outputs (e.g., schema:Dataset) and sitelinks to Wikipedia articles (schema:Article). Also used with the properties, name and description, and when mapping Wikidata to schema.org concepts[17].

[14] https://www.w3.org/TR/rdf11-concepts/
[15] https://www.w3.org/TR/rdf-schema/
[16] https://schema.org/
[17] Used as the object of the properties, equivalent class P1709, exact match P2888, super-property P2235, sub-property P2236, and equivalent property P1628

Prefix	Full URI	Semantics and Usage
skos:	`http://www.w3.org/2004/02/skos/core#`	Prefix of the W3C's Simple Knowledge Organization System (SKOS) namespace[18] used with the properties, prefLabel and altLabel
v:	`http://www.wikidata.org/value/`	Prefix for a value node in Linked Data's RDF. Note that the RDF Dump prefix is wdv:.
wd:	`http://www.wikidata.org/entity/`	Prefix for a subject node whose triples define the data for an entity (e.g., triples with the data for Douglas Adams Q42 are specified with the subject node, wd:Q42)
wdata:	`http://www.wikidata.org/wiki/Special:EntityData/`	Prefix for a subject node in an RDF Dump output, whose triples define the dump's metadata
wdno:	`http://www.wikidata.org/prop/novalue/`	Prefix for a class indicating that the property (the local name) has *no value* (different from an unknown value and discussed further under OWL usage)
wdref:	`http://www.wikidata.org/reference/`	Prefix for a provenance reference node in an RDF Dump. Note that the Linked Data prefix is ref:.
wds:	`http://www.wikidata.org/entity/statement`	Prefix for a statement node in an RDF Dump. Note that the Linked Data prefix is s:.
wdt:	`http://www.wikidata.org/prop/direct/`	Prefix for a property whose value has the "best" rank, provided without qualifiers or references

[18] https://www.w3.org/2004/02/skos/

Prefix	Full URI	Semantics and Usage
wdtn:	`http://www.wikidata.org/prop/direct-normalized/`	Prefix for a property whose value has the "best" rank, provided without qualifiers or references, and "normalized" (using SI units or full IRIs)
wikibase:	`http://wikiba.se/ontology#`	Prefix for the Wikidata meta-concepts (such as Item, Property, …)
xsd:	`http://www.w3.org/2001/XMLSchema#`	Prefix for the W3C XML Schema standard[19] used to identify datatype values (such as xsd:integer)

Table 2. SPARQL Prefixes and Their Semantics

While Figure 1 showed basic truthy and simple statement data for Christopher Armatas Q100049183, it is also valuable to review exemplary triples from more complex statement nodes, as well as reference and value nodes. Figure 2 illustrates the interplay of various item, property, statement, reference and value prefixes in the Linked Data output for Douglas Adams Q42.

[19] https://www.w3.org/XML/

Figure 2. Example of a Complex Statement Node with Qualifier, Value and Reference Nodes

A few additional clarifications are needed to understand the Wikidata Linked Data or Dump output in the context of RDF/OWL. These are addressed in the two subsections below.

Use of OWL in the Wikidata RDF

The most common occurrence of OWL in an RDF output is to specify that a Wikidata property predicate is an owl:DatatypeProperty (having a literal value) or an owl:ObjectProperty (having an IRI value). Some property predicates are defined as owl:DatatypeProperties – such as "date of birth" wdt:P569 or "ISBN 13-digit identifier" wdt:P212. Others, such as "stated in" wdt:P248 or "country of citizenship" wdt:P27, are defined as owl:ObjectProperties. Note, however, that all of the p:, :psv, :pqv or :prv predicates are defined as owl:ObjectProperties since they only ever reference a value node. (For more information on Wikidata property value types, see the subsection, Types of Wikidata Property Values.)

Other occurrences of OWL types and properties are related to indicating that a Wikidata property has *no possible value*. This declaration is a bit different than other "value"s in that it is defined as an instance of a class specified with OWL types and predicates, which is used in a multiple inheritance declaration. Figure 3 is an example of an instance which has no possible value (note that this is different than the value being unknown) for country of citizenship wdt:P27.

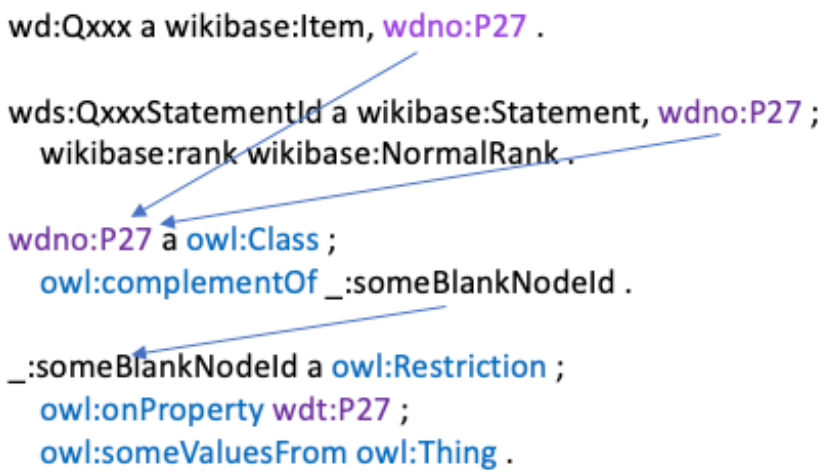


Figure 3. Example of an Instance with No Value for One of its Properties

The explanation of the semantics is straightforward, but the approach of using a class is very different than might be expected.

Another unexpected RDF encoding occurs to express that there is some (but unknown) value of a property. In this case, a blank node is referenced as the value of the property. This is shown in Figure 4 for the instance, Supercalifragilisticexpialidocious Q103, and the end time P582 of its You-Tube video.

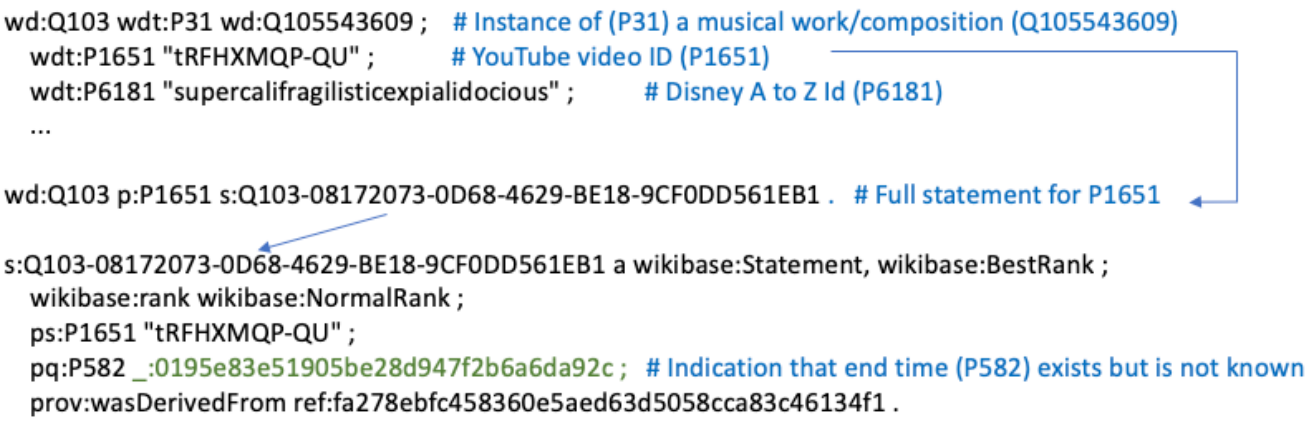


Figure 4. Example of the Acknowledgement that Some Value Exists for an Instance

Much more of OWL could be valuable if added to the RDF output (for example, the concept of disjointness for consistency analysis). This will be discussed in more detail in a future paper.

RDF Triples in Wikidata Linked Data, Dump and WDQS

Most of the conventions above hold true whether querying Wikidata using the WDQS endpoint, by utilizing the Linked Data output (for example, www.wikidata.org/entity/Q42.ttl), or by examining a downloaded RDF Dump output. However, there are a few differences:

- SPARQL prefixes are different:
  - The RDF Dump and WDQS "wdata:" prefix is "data:" for Linked Data
  - The RDF Dump and WDQS "wdref:" prefix is "ref:" for Linked Data
  - The RDF Dump and WDQS "wds:" prefix is "s:" for Linked Data
  - The RDF Dump and WDQS "wdv:" prefix is "v:" for Linked Data
- Wikibase-prefixed type definitions shown in Figures 1 and 2 (for example, "Wikibase:Item" in "wd:Q42 a Wikibase:Item") are not queriable in WDQS
  - An option would be to query based on the SPARQL prefix (for example, "wdref:" for reference nodes), but SPARQL FILTERing on the strings is not performant in WDQS
  - As an alternative to filter for:
    - Wikibase:Items - Query for "?wikibaseItem wikibase:sitelinks []" on WDQS or "?wikibaseItem ^schema:about/wikibase:sitelinks []" on non-WDQS endpoints, where sitelinks are references from the Wikidata Item to

another Wikimedia source, such as Wikipedia. Sitelinks are encoded in the RDF output as "?sourceIRI schema:about ?wikibaseItem" and are only used with Wikibase:Items.

- Wikibase:Statements - Query for "?wikibaseStatement wikibase:rank []". Ranks are encoded in the RDF output and are only used with Wikibase:Statements.
- Wikibase:References - Query for "?wikibaseRef ^prov:wasDerivedFrom []". Note that an InversePath RDF property path is used in the query. Provenance references are encoded in the RDF output from a Wikibase:Statement to a Reference (not using the InversePath). Only Wikibase:References are objects of the prov:wasDerivedFrom predicate.
- Wikibase:Properties - Query for "?wikibaseProp wikibase:propertyType []". Only Wikibase:Properties are objects of the wikibase:propertyType predicate.
- Lexeme-related entities such as a lexeme and its senses and forms - Query for the types, ontolex:LexicalEntry, ontolex:LexicalSense and ontolex:Form, respectively. A lexeme's sense is related using the predicate, ontolex:sense, and it's form is related using the predicate, ontolex:lexicalForm.

- Wdata: entity nodes (Linked Data's data:xx nodes) are not stored in the Blazegraph database backing WDQS. However, querying for various, specific metadata predicates is supported by using the pattern, "?s ?predicate ?o. VALUES ?predicate {schema:version schema:dateModified wikibase:statements wikibase:sitelinks wikibase:identifiers}". Both the RDF Linked Data and Dump/WDQS outputs are shown in Figure 5.
  - Note also that each item (wd:Qxxx) is defined as a type of schema:Dataset in the RDF Dump.
- As noted in both Figures 1 and 2, the value of the predicates, rdfs:label, skos:prefLabel and schema:name, is the same. Since that is true, only the value of rdfs:label is queriable in WDQS.
- The RDF output has two "special" values (no value and some value, as discussed in the OWL subsection, above). Since blank nodes are not identifiable/queriable, all references to "some value" (a value exists but is not known) are skolemized for the query service. Because of this, when querying to understand if a value exists but is unknown, the

syntax, "FILTER wikibase:isSomeValue(?someNode)" should be used (instead of FILTERing isBlank(?someNode). Using wikibase:isSomeValue returns a node with an IRI similar to www.wikidata.org/.well-known/genid/someHash.

```
RDF EntityData output:
  data:Q42 a schema:Dataset ;
    schema:about wd:Q42 ;
    cc:license <http://creativecommons.org/publicdomain/zero/1.0/> ;
    schema:softwareVersion "1.0.0" ;
    schema:version "2041528806"^^xsd:integer ;
    schema:dateModified "2023-12-31T16:17:39Z"^^xsd:dateTime ;
    wikibase:statements "309"^^xsd:integer ;
    wikibase:sitelinks "122"^^xsd:integer ;
    wikibase:identifiers "236"^^xsd:integer .
```

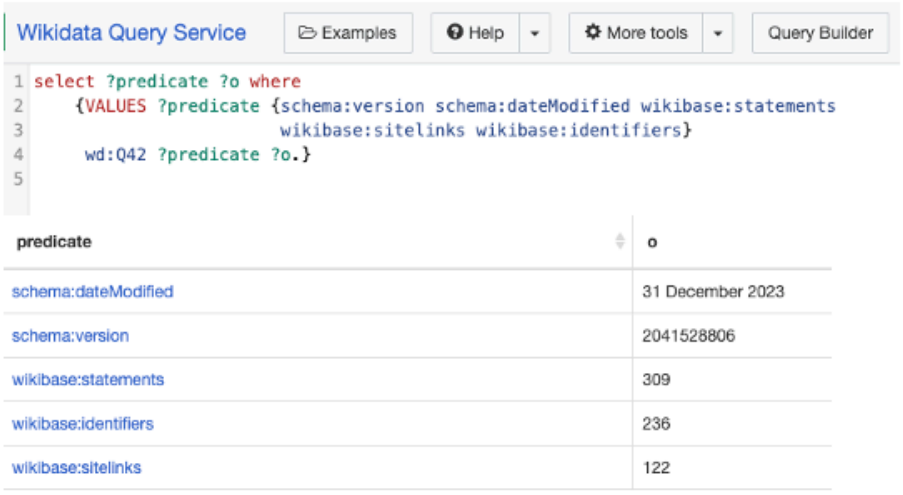


Figure 5. Metadata as Encoded for RDF Linked Data and for WDQS

Types of Wikidata Property Values

As mentioned in the OWL discussion above, different literal and IRI values are returned by the Wikidata properties. The current Wikidata property value types are defined on the Help:Data

Type web page[20]. Table 3 shows the correspondence between the Wikidata and RDF data types, with exemplary Wikidata values.

Wikidata Type	Example	RDF Data Type
WikibaseItem, WikibaseProperty, WikibaseLexeme, WikibaseSense, WikibaseForm	(Reference to a Wikidata entity such as wd:Q42 or wd:P269)	IRI/object reference
String (set of characters that do not need translation)	wd:Q42 wdt:P373 "Douglas Adams" (P373, Commons category)	xsd:string or "string_text"@lang_tag
Monolingual text (set of characters with a language tag where the string is specific to that language)	s:Q42-xxx pq:P2096 "Douglas Adams' gravestone"@en, "Grabstein von Douglas Adams"@de, "Douglas Adams'ın mezar taşı"@tr, ... (P2096, media legend)	"string_text"@lang_tag
External identifier (string identifier from a non-Wikibase system)	wd:Q42 wdt:P269 "026677636" (P269, IdRef ID)	xsd:string
External identifier, normalized value	wd:Q42 wdtn:P269 <http://www.idref.fr/026677636/id> (P269, IdRef ID)	IRI/object reference
URL	wd:Q42 wdt:P856 <https://douglasadams.com/> (P856, official website)	IRI/object reference

[20] The Help:Data Type web page defines most properties as "string-based". This is related to the Wikimedia storage and not the RDF Dump output. What is discussed in Table 3 is the RDF formatting.

Wikidata Type	Example	RDF Data Type
Commons media (reference to a file on Wikimedia Commons)	wd:Q42 wdt:P18 <http://commons.wikimedia.org/wiki/Special:FilePath/Douglas%20adams%20portrait%20cropped.jpg> (P18, image)	IRI/object reference
Commons geographic shape (reference to a map data file on Wikimedia Commons)	wd:Q1350565 wdt:P3896 <http://commons.wikimedia.org/data/main/Data:Germany.map> (Q1350565, Germany; P3896, geoshape)	IRI/object reference
Commons tabular data (reference to a tabular data file on Wikimedia Commons)	wd:Q61004609 wdt:P4150 <http://commons.wikimedia.org/data/main/Data:Weather.gc.ca/Monthly/8404201.tab>, <http://commons.wikimedia.org/data/main/Data:Weather.gc.ca/Almanac/8404201.tab> (Q61004609, Westbrook St Lawrence Canadian weather station; P4150, weather history)	IRI/object reference
Mathematical expression (string expressed in MathML syntax[21])	wd:Q47270 wdt:P4020 "$\n …\n$"^^<http://www.w3.org/1998/Math/MathML> (Q47270, half-life; P4020, ISQ dimension)	xsd:string

[21] https://en.wikipedia.org/wiki/MathML

Wikidata Type	Example	RDF Data Type
Musical notation (string expressed in LilyPond notation[22])	wd:Q1635257 wdt:P6686 "\\relative c''{\\key f \\major \\set Staff.midiInstrument=#\"violin\"\\tempo\ "Allegro moderato Très doux\"4= 120a4\\p(g8a)e4(d)a'8(f c'e)d2}" (Q1635257, Ravel's String Quartet; P6686, musical motif)	xsd:string
Quantity (decimal value with qualifiers, amount, unit and uncertainty bounds)	(See Figure 6)	xsd:decimal
Time	(See Figure 7)	xsd:dateTime
Globe coordinate (geographical position as longitude-latitude on a "globe")	(See Figure 2, where the "globe" is Earth Q2)	"Point(long lat)" ^^geo:wktLiteral (GeoSPARQL data type)

Table 3. Correspondence of Wikidata and RDF Data Types

[22] https://en.wikipedia.org/wiki/LilyPond

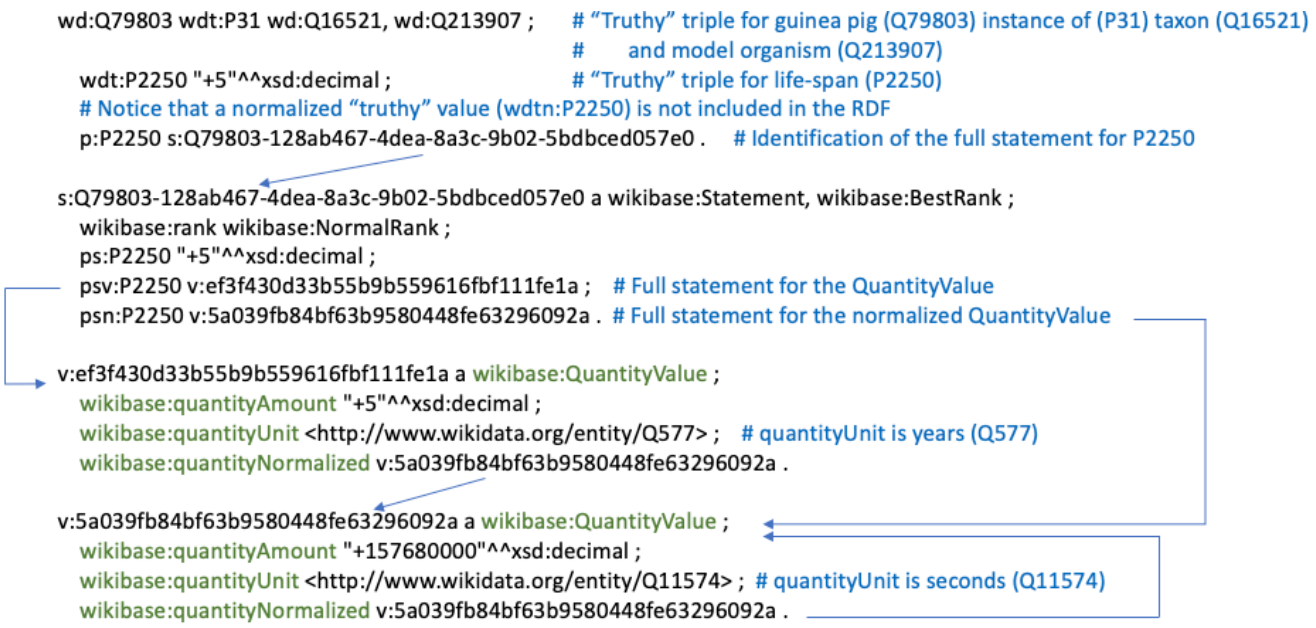


Figure 6. Example of Wikidata Quantity Property Value

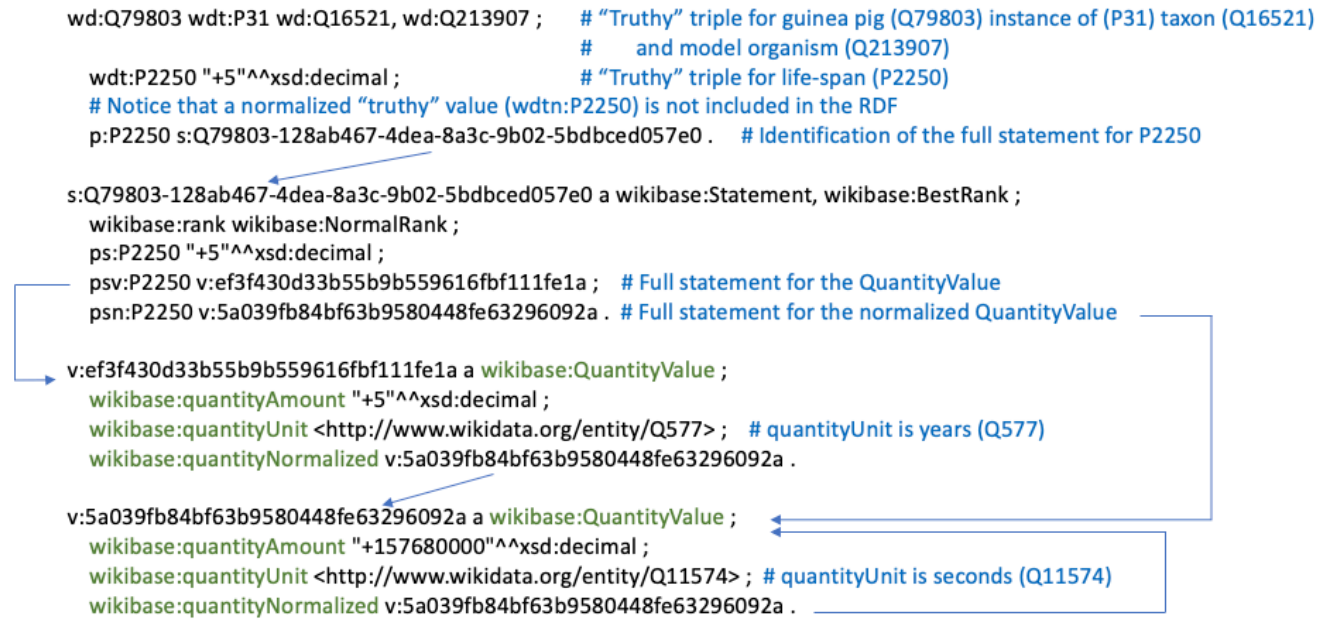


Figure 7. Example of Wikidata Time Property Value

The Wikidata data type of a property can be determined from the RDF output by searching/querying for the value of the property's wikibase:propertyType. That is a reference to one of the entries in the first column of Table 3.

In addition to the Wikidata properties, other "standard" properties are included in the RDF output. These were mentioned (related to the namespaces where they are defined) in Table 2. Their values are all as expected.

Mapping of Wikidata to RDF/OWL

Beyond the Dump output's use of RDF/OWL and a mapping of the Wikidata data value types to XSD/OWL/GeoSPARQL, a bit more detail is needed to fully map Wikidata to RDF/OWL. To capture this information, Table 4 expands the discussion to include the correspondence between Wikidata and OWL2 standard classes and properties.

Wikidata Concept	RDF/OWL Mapping (Turtle syntax)
Item	owl:Class or owl:NamedIndividual
Property	owl:DatatypeProperty or owl:ObjectProperty
Statement and Qualifier	(use of) RDF* (RDF-star)
P31, instance of	rdf:type
P279, subclass of	rdfs:subClassOf
P1647, subproperty of **P2235, external superproperty** **P2236, external subproperty**	rdfs:subPropertyOf (or its inverse to indicate superproperty)

Wikidata Concept	RDF/OWL Mapping (Turtle syntax)
P2302, property constraints	rdfs:domain and rdfs:range owl:onProperty and owl:onDatatype restrictions[23] owl:FunctionalProperty and owl:InverseFunctionalProperty owl:SymmetricProperty
P1545, series ordinal (references a statement)	rdf:first and rdf:rest
P361, part of **P527, has part**	rdfs:member (and its inverse)
P1709, equivalent class	owl:equivalentClass
P1628, equivalent property	owl:equivalentProperty
P1696, inverse property	owl:inverseOf
P1889, different from	owl:differentFrom
P2737, union of	owl:unionOf (where the list is provided as a "list of values as qualifiers" Q23766486)
P2738, disjoint union of	owl:disjointUnionOf (where the list is provided as a "list of values as qualifiers" Q23766486)
P2888, exact match	owl:equivalentClass (for classes) owl:sameAs (for individuals)

[23] Restrictions and the related owl predicates are discussed further in the section, Constraints and Quality Control

Wikidata Concept	RDF/OWL Mapping (Turtle syntax)
P460, same as (similar to P2888 but may be disputed)	owl:sameAs
P1889, different from	owl:AllDifferent (similar semantic)
P1552, has characteristic, referencing irreflexivity (Q54933368) or reflexivity (Q54933018)	owl:ReflexiveProperty, owl:IrreflexiveProperty
In a few cases modeled as distinct properties (e.g., has part P527 and does not have part P3113)	owl:NegativePropertyAssertion
N/A	owl:intersectionOf, owl:complementOf owl:propertyDisjointWith owl:propertyChainAxiom rdf:AsymmetricProperty owl:TransitiveProperty

Table 4. Correspondence between Wikidata and OWL Concepts

All of the above information can be combined to create a Wikidata RDF/OWL encoding, which is work currently under development.

Wikidata Ontologies

There are several "ontologies" in Wikidata and mappings to external ontologies/schemas such as the Basic Formal Ontology (BFO) and schema.org.

Within Wikidata, one could consider its data model (discussed in the previous section) as a meta-ontology. In fact, the Wikibase Data Model web page states that the "data model provides a metamodel or ontology for describing real world entities" (downloadable as ontology-1.0.owl).

Alternately, one could view Wikidata's top-level concepts as forming an ontology. Top-level concepts could be identified by querying for all Wikidata items that are themselves superclasses but are not a "subclass of" (P279) another item (see Figure 8 for the query and results obtained from the QLever Wikidata SPARQL endpoint[24]). Regrettably, there are over 10,000 results – many of which do not reasonably qualify as "top-level".

[24] That query timed out on WDQS.

```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?s (COUNT(distinct ?other) as ?cnt) WHERE
{?other wdt:P279 ?s . MINUS {?s wdt:P279 ?superclass}
 FILTER (!isBlank(?s))}
GROUP BY ?s ORDER BY DESC(?cnt)
```

Query results:

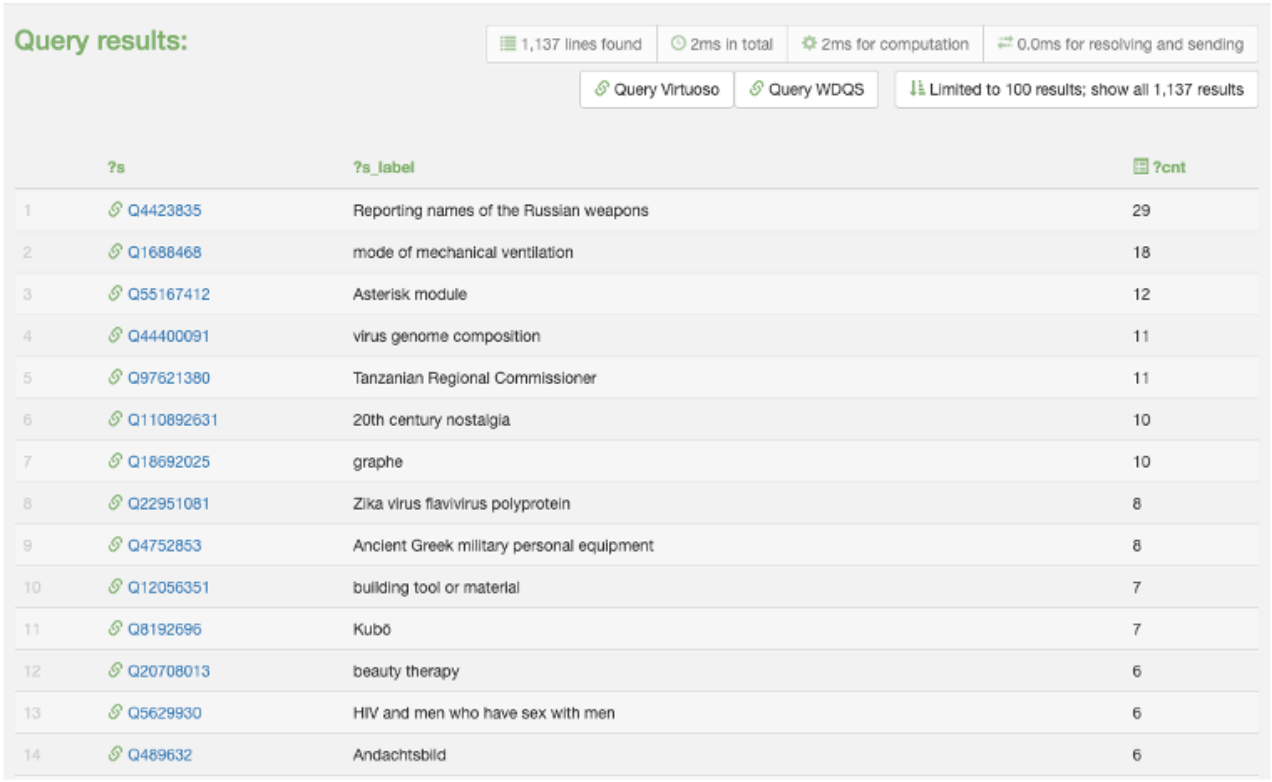
10,484 lines found | 2ms in total | 1ms for computation | 1ms for resolving and sending

Query Virtuoso | Query WDQS | Limited to 100 results; show all 10,484 results

	?s	?s_label	?cnt
1	Q9592	Catholic Church	234
2	Q6451898	Girl Guides	180
3	Q83140574	SH3 domain, protein family	180
4	Q7543599	Small ubiquitin like modifier 1	150
5	Q12271	architecture	134
6	Q30046	Vitis vinifera	105
7	Q5009781	fms related receptor tyrosine kinase 3	105
8	Q2621182	minister of culture	103
9	Q3457052	Road network in France	101
10	Q4115952	commerce minister	98
11	Q1449973	environment minister	97
12	Q12303322	minister of labor	94
13	Q21111827	ubiquitin C	81
14	Q1857	lutetium	81
15	Q27644149	list of members of the Leopoldina – National Aca[...]	77

Figure 8. Wikidata Query of Superclasses that Have No Superclass Themselves

```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?s (COUNT(distinct ?other) as ?cnt) WHERE
{?other wdt:P279 ?s . MINUS {?s wdt:P279 ?superclass}
 MINUS {?s wdt:P31 ?instanceOf} FILTER (!isBlank(?s))}
GROUP BY ?s ORDER BY DESC(?cnt)
```



Query results:

1,137 lines found | 2ms in total | 2ms for computation | 0.0ms for resolving and sending

Query Virtuoso | Query WDQS | Limited to 100 results; show all 1,137 results

	?s	?s_label	?cnt
1	Q4423835	Reporting names of the Russian weapons	29
2	Q1688468	mode of mechanical ventilation	18
3	Q55167412	Asterisk module	12
4	Q44400091	virus genome composition	11
5	Q97621380	Tanzanian Regional Commissioner	11
6	Q110892631	20th century nostalgia	10
7	Q18692025	graphe	10
8	Q22951081	Zika virus flavivirus polyprotein	8
9	Q4752853	Ancient Greek military personal equipment	8
10	Q12056351	building tool or material	7
11	Q8192696	Kubō	7
12	Q20708013	beauty therapy	6
13	Q5629930	HIV and men who have sex with men	6
14	Q489632	Andachtsbild	6

Figure 9. Wikidata Query Similar to Fig 8 But Removing Items That Are Also Instances

Figure 9 shows the query and top results when removing any superclasses that are "instance(s) of" (P31) or "subclass(es) of" (P279) another item. This reduces the number of results to over 1000 but similarly, these do not qualify as "top-level".

Another alternative is to consider the items identified by the Ontology WikiProject. These are "class" Q16889133 (semantically equivalent to owl:Class, a means to group items having similar characteristics) and "entity" Q35120 (similar to owl:Thing, in that it can take the role of a superclass or be the type of an instance). And, similar to owl:Thing, all Wikidata "class"es are subclasses of "entity" (via the intermediary subclasses, "arbitrary entity" Q7048977 and "collective entity" Q99527517).

Working against this approach, the top-level tree under "entity" Q35120 is confusing and arbitrary[25]. For example, there are 38 direct subclasses of "entity" including two items both labelled as "meaning" (Q16877777 and Q16877783). Also included are linguistic thematic concepts such as "object" (Q488383), "recipient" (Q119515609) and "result" (Q2995644), overlapping concepts such as "location" (Q115095765) and "origin" (Q3885844), and concepts based on perspective such as "current entity" Q96196524 (existing "in the present") and "former entity" Q15893266.

Further down the subclassing tree, one finds items such as "Mother Nature" (Q1402540), which is an instance of a "personification" (Q207174), "literary character" (Q3658341) and "theatrical character" (Q3375722)[26], but is also a subclass of (as opposed to an instance of) an "abstract entity" (Q7048977). The latter makes "Mother Nature" a peer to the top level "class" Q16889133, which seems unreasonable.

Several reasons for these types of confusion are the lack of clear modeling guidelines, constraints and oversight in Wikidata, the desire to capture the structured data for *each* Wikipedia entry (even if those concepts are very similar semantically - e.g., leading to the creation of the two "meaning" subclasses of "entity"), and problems caused by correcting a subclassing/instantiation tree which in turn affects the integrity of other, seemingly unrelated items.

Given these issues and more[27], it is valuable to instead focus on understanding and working with specific Wikidata's instances and inheritance hierarchies (a more bottom-up approach consistent with the overall design and evolution of Wikidata), and/or to examine and supplement mappings of Wikidata entities to curated ontologies (such as schema.org). The latter approach would aid in creating a stable backing ontology for applications and reasoning. These topics are discussed in more detail in the following subsections.

[25] A complete traversal of the "entity" subclassing tree can be seen on the web page, https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology/Top-level_ontology_list

[26] Note also that both literary and theatrical character as instances of "types of fictional character" and subclasses of "fictional character".

[27] https://commons.wikimedia.org/wiki/File:Wikidata_ontology_issues_%E2%80%94_suggestions_for_prioritisation_2023.pdf

Frequently Used Wikidata Classes and Properties

One approach to understanding Wikidata is to compile a list of subject areas where there are many defined items (classes and instances). Such a list could be derived from the general topics of active WikiProjects, Wikipedia's content areas and frequently occurring items (the latter are shown in Figure 10, based on the RDF Dump from the end of 2021).

The list of oft-used subject areas is summarized as follows:

- "Science" Q336, "branch of science" Q2465832 (a subclass of Q336/science) and "taxon"/taxonomic study Q16521 (ranked fifth in Figure 10)
    - Note that "mathematics" Q395, "society" Q8425 and "technology" Q11016 (referenced as Wikipedia Content) are "subclasses of" P279 or "studied by" P2579 Q336/science
    - Note that "gene" Q7187 (ranked eighth in Figure 10) is "part of" P361 "genome" Q7020 which is "studied by" P2579 "genetics" Q7162, a subclass of Q336/science, that "chemical compound" Q11173 (ranked tenth in Figure 10) is an "instance of" a "group or class of chemical substances" Q17339814, which is "studied by" P2579 "chemistry" Q2329, and that "protein" Q8054 (ranked eleventh in Figure 10) is an "instance of" P31 of "structural class of chemical entities" Q47154513 (8, 10, 11) which is a "subclass of" a "group or class of chemical substances" Q17339814, which is "studied by" P2579 "chemistry" Q2329
- "Philosophy" Q5891 and "branch of philosophy" Q22811234 (a "part of" P361 Q5891/philosophy)
- "History" Q309, "study of history" Q1066186, including events/"occurrence" Q1190554, conditions/"state" Q3505845 and "process" Q3249551 (a series of events)
- "Culture" Q11042, "heritage" Q2434238 (which includes "cultural heritage" Q210272, a "part of" P361 Q11042/culture)
    - Note that "religion" Q9174 is a subclass of "belief system" Q5390013, which is a "part of" P361 Q11042/culture
- "Work" Q386724, "creative work" Q17537576 (a subclass of Q386724/work) and "GLAM"-related Q1030034 (galleries, libraries, archives and museums) information (a "facet of" P1269 Q210272/cultural heritage)

    - Note that "scholarly articles" Q13442814 (ranked first in Figure 10) is a subclass of "work" Q386724, as is "painting" Q3305213 (ranked twelfth), "film" Q11424 (ranked fourteenth), "encyclopedia article" Q13433827 (ranked seventeenth) and "collection" Q2558072 (ranked nineteenth)
- "Geographic entity" Q27096213 and "space object" (Q4235019)
    - Note that "astronomical object" Q6999 (ranked second in Figure 10) is a subclass of Q4235019/space object
    - Note also that "village" in China Q13100073 (ranked thirteenth in Figure 10), "human settlement" Q486972 (ranked fifteenth), "mountain" Q8502 (ranked eighteenth) and "street" Q79007 (ranked twentieth) are subclasses of Q27096213/geographic entity
- "Agent" Q24229398 (with the subclass, "person or organization" Q106559804), "role" Q4897819, "group of living things" Q16334298 (with the subclass, "group of humans" Q16334295), "behavior" Q9332 (with "has part" P527 "human behavior" Q3769299 and "animal behavior" Q2990593), and "activity" Q1914636
    - Note that "human" Q5 (ranked third in Figure 10) is a subclass of Q106559804/person or organization) and is related via "has characteristic" P1552 to full name Q1071027 which "has part" P527 "family name" Q101352 (ranked sixth in Figure 10)
- Other subject areas related to "human"s Q5 and agents:
    - "Nutriment" Q1422299 (such as foods and beverages)
    - "Physiological condition" Q7189713, including its subclasses, "health" (Q12147) and "disease" Q12136
    - Laws, regulations, etc. which are subclasses of "rule" Q1151067 and "legal norm" Q216200
- "Resource" Q1554231 (such as economic, financial or information resources), "service" Q7406919 and "goods" Q28877

Rank	Subgraph	Subgraph Name	Number of items	% of WD items	
1	Q13442814	scholarly article	37,362,641	39.75	6
2	Q6999	astronomical object	8,412,914	8.95	1
3	Q5	human	9,315,444	9.91	9
4	Q4167836	Wikimedia category	4,840,195	5.15	7
5	Q16521	taxon	3,180,248	3.38	3
6	Q101352	family name	481,445	0.51	1
7	Q4167410	Wikimedia disambiguation page	1,359,804	1.45	1
8	Q7187	gene	1,196,361	1.27	1
9	Q11266439	Wikimedia template	845,852	0.9	1
10	Q11173	chemical compound	1,223,387	1.3	9
11	Q8054	protein	986,599	1.05	8
12	Q3305213	painting	539,468	0.57	5
13	Q13100073	village-level division in China	588,477	0.63	5
14	Q11424	film	263,070	0.28	4
15	Q486972	human settlement	563,958	0.6	3
16	Q13406463	Wikimedia list article	334,939	0.36	3
17	Q13433827	encyclopedia article	512,141	0.55	3
18	Q8502	mountain	525,553	0.56	3
19	Q2668072	collection	500,968	0.53	3
20	Q79007	street	578,926	0.62	3

Figure 10. Top 20 Types of the Most Frequently Referenced Wikidata Items

As regards properties, a list of the most used properties is also important to highlight. The top 20 of these properties (as of the end of January 2024) are shown in Figure 11.

Property	Quantity of item pages
cites work (P2860)	292467905
series ordinal (P1545)	175532936
author name string (P2093)	138579750
instance of (P31)	114429000
retrieved (P813)	97568673
stated in (P248)	97373672
reference URL (P854)	74284349
PubMed ID (P698)	64671497
title (P1476)	50577671
publication date (P577)	50005238
published in (P1433)	42240558
page(s) (P304)	38195852
volume (P478)	37469394
issue (P433)	35637817
apparent magnitude (P1215)	33143984
astronomical filter (P1227)	33143941
DOI (P356)	32397102
author (P50)	31139791
catalog code (P528)	28922366
main subject (P921)	28825181

Figure 11. Top 20 Frequently Used Wikidata Properties

The classes and properties in the above lists are examined in more detail related to external ontologies (directly below).

Alignment with External Ontologies

In order to determine mappings from Wikidata to other ontologies, the queries shown below were executed. Due to timeout issues on WDQS, the queries were executed on the QLever SPARQL endpoint.

- Query for all properties related to external references

```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wikibase: <http://wikiba.se/ontology#>
SELECT DISTINCT ?prop ?extPred ?ext WHERE {
   VALUES ?extPred {wdt:P2235 wdt:P2236 wdt:P1628}
   ?prop ?extPred ?ext .
   FILTER (!CONTAINS(str(?ext),"http://www.wikidata.org")) .
}
```

- Query for external references that are equivalent classes, exact matches or narrower

```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
SELECT DISTINCT ?item ?extPred ?ext WHERE {
   VALUES ?extPred {wdt:P1709 wdt:P2888 wdt:P3950}
    ?item ?extPred ?ext .
    { {?item wdt:P279 ?x} UNION {?y wdt:P279 ?item} } # Items in class hierarchy
    FILTER (!CONTAINS(str(?ext),"http://www.wikidata.org")) .
}
```

Note that the last query checks for subclass hierarchies using P279 (subclass of). This was done to avoid the many millions of "exact match" references due to matching instances. The latter skews the results related to investigating ontology definitions (Tboxes[28]).

To do this processing, a Jupyter notebook was created to total the references (across http, https and purl.org URLs). The notebook Wikidata-Count-Ext-Links.ipynb is available in the Wikidata-and-OWL GitHub repository, in the *notebooks* subdirectory.

Links with more than 50 references are shown below, with some details about their available information and counts:

- identifiers.org, 199777 references, providing resolvable references for life sciences data
- OBOFoundry/purl.obolibrary.org, 57677 references, providing resolvable references for OBO-related PURLs (permanent URLs)
- orpha.net, 8573 references, providing data on rare diseases with backing ontology (Orphanet Rare Disease Ontology)

[28] https://en.wikipedia.org/wiki/Abox

- publications.europa.eu, 5845 references, providing information on EU laws, institutions, and more related to managed publications, with backing vocabularies, schemas and ontologies
- rhea-db.org, 4392 references, providing information on chemical reactions, with backing ChEBi ontology
- schema.org, 791 references, providing a vocabulary for creating structured data on the Web related to a wide variety of resources
- wordnet-rdf.princeton.edu, 543 references, providing WordNet-based lexical information and an ontology
- UniProt/purl.uniprot.org/www.uniprot.org, 501 references, providing protein sequencing and functional information with a core ontology and SPARQL endpoint
- tcdb.org, 494 references, providing a database and classification system for membrane transport proteins (based on the TC-System classification system)
- DBpedia/dbpedia.org, 424 references, providing a cross-domain, general-purpose knowledge graph and ontology (DBO, DBPedia ontology)
- ncbi.nlm.nih.gov, 198 references, providing biotechnology data with a backing taxonomy
- w3.org, 161 references, defining the RDF, RDFS, OWL, SKOS, time, vCard, organization and activity streams standards and ontologies
- lexinfo.net, 141 references, providing data categories for modeling lexicons and dictionaries as defined by a backing ontology
- id.loc.gov, 125 references, providing access to ontologies, vocabularies and other information for bibliographic data
- pcp-on-web.de, 92 references, providing the vocabulary for Early Modern Career Patterns
- purl.org/ontology, 79 references (to the bibliographic, music, BBC programs and BBC wildlife ontologies)
- purl.org/dc, 78 references, defining meta-data, educational resource and bibliographic specifications and ontologies, created by the Dublin Core Metadata Initiative
- purl.org/spar, 71 references, defining the Semantic Publishing and Referencing (SPAR) Ontologies which consist of multiple OWL ontologies
- cv.iptc.org, 64 references, providing a consolidated vocabulary for news information, created by the International Press Telecommunications Council

- d-nb.info, 56 references, defining Gemeinsame Normdatei's GND ontology for name disambiguation in library data
- purl.org/coar, 56 references, defining a set of frameworks, standards and vocabularies for repository information, created by the Confederation of Open Access Repositories (COAR)

Many of the above references are related to specialized efforts in the domains of biochemistry and medicine, research and publishing, and lexicology. There are two external references which address a broad range of domains (schema.org and DBpedia). In addition, the W3C standards related to people and organizations (vCard and organization) are valuable since they are intended as generic models.

As part of on-going work, the four ontologies (schema.org, DBpedia, W3C vCard and W3C organization) are coupled with constraint definitions (as described in the next section) to create an RDF/OWL "top-level" ontology. Details of that work will be published later in 2024.

Wikidata Constraints and Quality Control

Wikidata constraints are associated with a property definition. The specific constraints associated with a property can be queried in the RDF Dump by searching on "?property wdt:P2302 ?constraint" or "?property wdt:P1552 wd:Q100884525" (where P2302 indicates a property constraint but that identifier is not used for complex constraint values label Q100884525). The constraint that is referenced defines specific semantics – such as the format constraint Q21502404 defining the format of the property value.

Although the applicability of a constraint to a property is defined in the RDF Dump, an item's *violation* of the constraint is NOT reported. This is visible on the individual wiki pages for an entity, or as noted on the Wikidata Database Reports Constraint Violations pages (which are updated daily). Figure 12 is an example of a constraint reported on an individual item wiki page, while Figure 13 shows specific *mandatory* constraints defined in Wikidata and the items that violate them. The figure was generated from the 13 February 2024 report which listed almost 2300 unique constraints

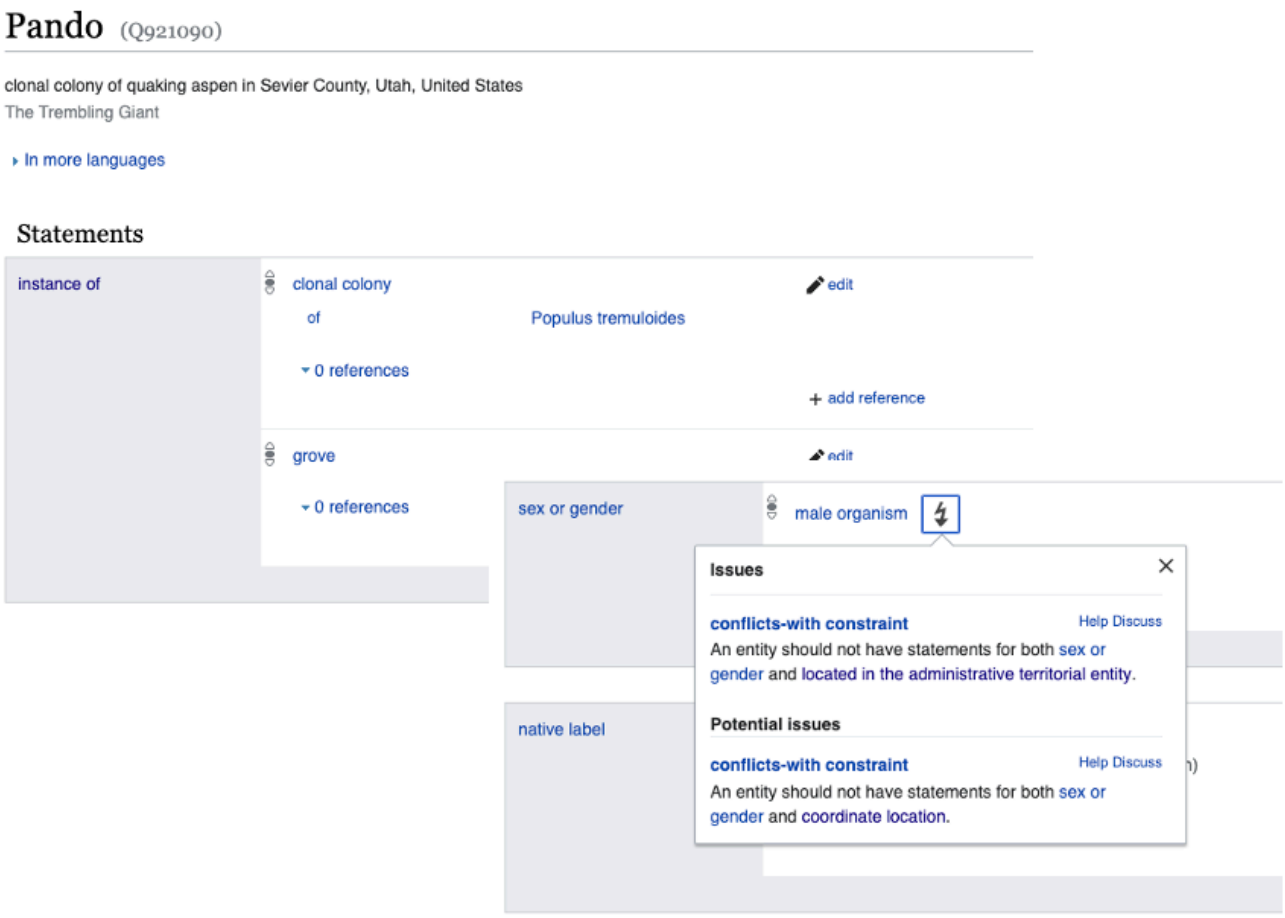


Figure 12. Constraint Violation Example Displayed on HTML Wiki Page

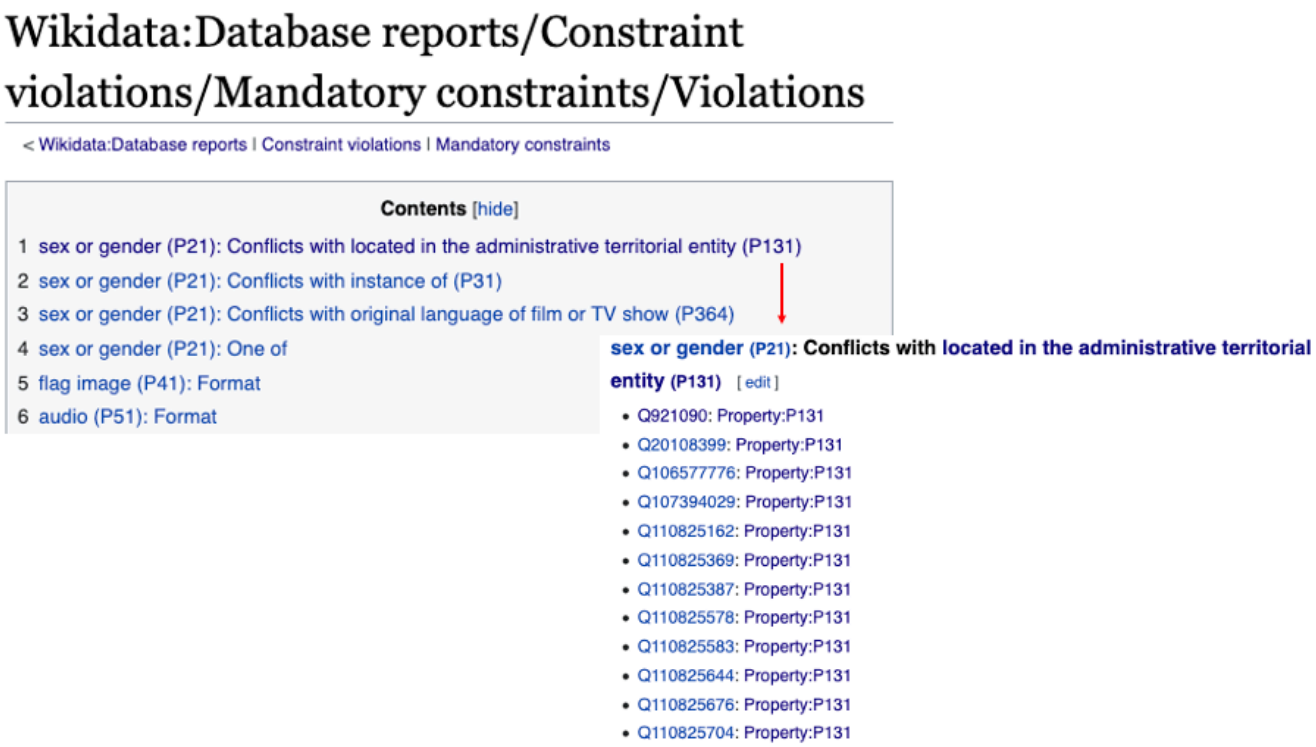


Figure 13. Data in the Wikidata Database Report of Mandatory Constraint Violations

Constraint violations do not necessarily indicate errors in the data. The violation could be an error, or indicate that the constraint is incorrectly defined or that an exception to the constraint is needed but has not yet been recorded[29]. A conservative approach to handling constraint violations would be to remove the triples of any items which are violations of any properties of interest. However, this approach may discard valid information along with errors.

A necessary question involves discovering and hopefully resolving constraint violations of interest. One means of discovery is shown in the Jupyter notebook, Wikidata-Constraint-Violations.ipynb (on GitHub). The approach starts with retrieving the mandatory constraint violations web page (using the Python *requests* module), parsing it (using the *html.parser* module), and assembling the details. The latter are output in an RDF Turtle file using the special prefix "wikidata-owl:" which is defined as <urn:wikidata-owl:> with additional

[29] For example, the character Leela from Futurama (Q121841) has a date of birth P569 that is in the future, which is allowed since she is a fictional character.

information defined using RDF-star (which is supported by many graph databases which input RDF).

Specifically, a violation would be output as a triple of the form:

<itemIRI> wikidata-owl:violatedProperty {
Optional_other_property_or_value ;
wikidata-owl:violationType <violationTypeIRI> ;
wikidata-owl:constraintText "stringTextFromViolationsReport"
} <propertyIRI>

Where the "*optional_other_property_or_value*" is defined based on the violation type and detailed in the bullet list below. For example, for the violation shown in Figure 12, the Turtle output would be:

wd:Q921090 wikidata-owl:violatedProperty {
wikidata-owl:conflictingProperty wd:P131 ;
wikidata-owl:violationType wd:Q21502838 ;
wikidata-owl:constraintText "sex or gender (P21): Conflicts with located in the administrative territorial entity (P131)"
} wd:P21

At a minimum, the existence of the statement would indicate reduced confidence in the information conveyed by the specified property for the indicated item.

At this time, Wikidata-Constraint-Violations.ipynb outputs violation triples for the following constraint types:

- Allowed-entity-type constraint (Q52004125); Violations indicate that the property should not be used with/is invalid for the entity
  - There is no *optional_other_property_or_value*
- Conflicts-with constraint (Q21502838); Violations indicate that the semantics of two different properties are not logical/not consistent for the item, and therefore the properties should not be used together
  - The "conflicting" property is referenced by the predicate, wikidata-owl:conflictingProperty

- Contemporary constraint (Q25796498); Violations indicate that the subject and object of the property triple should coexist at some point in time but do not
    - The other entity which is not contemporary is referenced by the predicate, wikidata-owl:nonContemporaryWith
- Format constraint (Q21502404); Violations indicate that there is a formatting error in the property value
    - There is no *optional_other_property_or_value*
    - Note that the format (defined as a regular expression) is referenced by the predicate, pq:P1793, in the property's constraint definition
- Integer constraint (Q52848401); Violations indicate that the property value should be an integer but is not
    - There is no *optional_other_property_or_value*
- Inverse constraint (Q21510855); Violations indicate that there is only a triple defined relating the subject to the object, but a corresponding triple in the reverse direction should exist
    - The other entity which should have an inverse relationship is referenced by the predicate, wikidata-owl:missedInverse
    - Note that the property defined as the "inverse" is referenced by the predicate, pq:P2306, in the original property's constraint definition
- None-of constraint (Q52558054); Violations indicate that the property value is erroneous, a better alternative exists or should not be used for other reasons
    - The invalid property value is specified as wikidata-owl:invalidItem
- One-of constraint (Q21510859); Violations indicate that the property value (the object of the property triple) is not defined as an item from a predefined set, and therefore is erroneous
    - The entity that is referenced by the property but has an invalid value is specified using the predicate, wikidata-owl:invalidItem
    - Note that if wikidata-owl:invalidItem is not specified, then the referenced entity has no instance of or subclass of declaration
    - Note that the allowed item values of the referenced entity are specified using the predicate, pq:P2305, in the original property's constraint definition
- Referenced-property-not-one-of constraint (aka, 'item requires statement' constraint, Q21503247); Violations indicate that the item using the property does not itself declare

a triple with another, specific predicate, or that the triple is declared but its property value is not one of a predefined set (as an example of the latter, if an item's Google Knowledge Graph ID (P2671) value is a string and not a graph reference, then the item would carry this violation)
    - The missing property declaration or declaration with an invalid item value is specified using the predicate, wikidata-owl:missingOrInvalidProperty
    - The allowed item value(s) for the wikidata-owl:missingOrInvalidProperty are provided using the predicate, wikidata-owl:allowedValue
    - The entity that is referenced by the property but has an invalid value is specified using the predicate, wikidata-owl:invalidItem
    - Note that if wikidata-owl:invalidItem is not specified, then the referenced entity has no instance of or subclass of declaration
- Single-value constraint (Q19474404); Violations indicate that more than one value is defined for a property that should be single-valued
    - There is no *optional_other_property_or_value*
- Subject-type constraint (Q21503250); Violations indicate that the referencing entity (the subject of the property triple) is not a subclass or instance of one of the required types
    - The allowed item value(s) are referenced by the predicate, wikidata-owl:allowedValue
    - The current, invalid type(s) for the item are referenced by the predicate, wikidata-owl:invalidItem
    - Note that if wikidata-owl:invalidItem is not specified, then no subclass or instance of declaration was provided for the entity
- Symmetric constraint (Q2510862); Violations indicate that the subject/object of the property triple should also be defined reversing the order (e.g., object – property – subject), but is not
    - The other entity which is missing a triple relating it to the original subject is referenced by the predicate, wikidata-owl:missedSymmetric
- Unique-value constraint (Q21502410); Violations indicate that the combination of the property and value should be relevant for only one item, but is used with multiples
    - The duplicated value is given by the predicate, wikidata-owl:duplicatedValue, or if the value is a string, by the predicated, wikidata-owl:duplicatedStringValue

    - A violation triple is defined for each item that uses the property-value pair, which enables querying for all entities with that value
- Value-type constraint (Q21510865); Violations indicate that the referenced entity (the object of the property triple) is not a subclass or instance of one of the required types
    - The allowed item value(s) are referenced by the predicate, wikidata-owl:allowedValue
    - The invalid referenced object is identified using the predicate, wikidata-owl:invalidItem

A next step would be to proactively use the information encoded in the violation triples to selectively delete triples from the RDF Dump (such as properties with erroneous values), or to insert triples (for example, adding inverse relationships where they are missing). These types of changes could be accomplished using SPARQL DELETE/INSERT statements issued to the graph database where the RDF Dump (or subset of the dump) is stored. In addition, this information could be used by a bot to correct the backing triples in Wikidata.

All of the above are elements of on-going work related to creating an RDF/OWL Wikidata encoding.