Wikidata: A large-scale collaborative ontological medical database

Houcemeddine Turki[1], Thomas Shafee[2], Mohamed Ali Hadj Taieb[3], Mohamed Ben Aouicha[4], Denny Vrandečić[5], Diptanshu Das[6], Helmi Hamdi[7]

[1]Faculty of Medicine of Sfax, University of Sfax, Sfax, Tunisia
[1]turkiabdelwaheb@hotmail.fr
[2]La Trobe University, Bundoora, Melbourne, Victoria, Australia
[2]T.Shafee@LaTrobe.edu.au
[3,4]Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia
[3]mohamedali.hadjtaieb@gmail.com [4]mohamed.benaouicha@fss.usf.tn
[5]Google LLC, Mountain View, California, United States of America
[5]vrandecic@google.com
[6]Institute of Neurosciences Kolkata (I-NK), Kolkata, India
[6]das.diptanshu@gmail.com
[7]CUFE, Université de Sherbrooke, Sherbrooke, Canada
[7]helmi.hamdi@usherbrooke.ca

Highlights:

- Wikidata is a flexible biomedical semantic knowledge-base
- It has the potential to perform a much better role as a large-scale biomedical semantic resource
- Tools are available to query, enrich, verify and process the biomedical knowledge

Abstract: Created in October 2012, Wikidata is a large-scale, human-readable, machine-readable, multilingual, multidisciplinary, centralized, editable, structured, and linked knowledge-base with an increasing diversity of use cases. Here, we raise awareness of the potential use of Wikidata as a useful resource for biomedical data integration and semantic interoperability between biomedical computer systems. We show the data model and characteristics of Wikidata and explain how this database can be automatically processed by users as well as by computer methods and programs. Then, we give an overview of the medical entities and relations provided by the database and how they can be useful for various medical purposes such as clinical decision support
Keywords: Wikidata, Biomedical ontology, Semantic resources, Multilingual resources, Medical databases

Introduction

With the rising use of artificial intelligence in medicine, there is a clear need to develop a large-scale, multi-lingual, machine-readable, human-readable, editable, structured and linked representation of all medical knowledge for processing biomedical data by computational methods (Patel, et al., 2009). One way to achieve this is the use of a unified representation such as a biomedical semantic knowledge database. It provides an exhaustive, yet ever-evolving, list of biomedical concepts such as genes, diseases, drugs, symptoms and their names in multiple languages (Collier, et al., 2006). These concepts are linked by taxonomic relations, such as “instance of” and “subclass of”, to allow the data to be classified, categorized and indexed (Smith, et al., 2005). Non-taxonomic associative relations (e.g. “drug used for treatment” and “risk factor”) should create a further link between medical concepts to give biomedical computer programs the ability to clinically analyze provided inputs (such as Electronic Health Records, IoT data, and X-ray images) as well as the ability of returning needed outputs (Bodenreider, 2008). In recent decades, many large-scale taxonomies including SNOMED-CT (Systematized Nomenclature of Medicine – Clinical Terms), MeSH (Medical Subject Headings), UMLS (Unified Medical Language System), and FMA (Foundational Model of Anatomy) (Bodenreider, 2008; Ivanović & Budimac, 2014) and multiple ontologies involving non-taxonomic biomedical associations[1]

[1]E.g. drug-target relations (Sun, 2015), drug-disease relations (Sun, 2015), disease-gene relations (Sun, 2015), and disease-disease risk relations (Xu, Li, & Wang, 2014).

(Sun, 2015; Xu, Li, & Wang, 2014) have been developed for a variety of biomedical applications like biomedical question answering (Athenikos & Han, 2010), clinical decision support (De Potter, et al., 2012) and the automation of computed tomography procedures (De Silva, MacDonald, Paterson, Sikdar, & Cochrane, 2011). Some of these resources such as BabelMeSH and SNOMED-CT are currently available in multiple languages and can be used to process biomedical data provided in languages other than English (Liu, Fontelo, & Ackerman, 2006; Henriksson, Skeppstedt, Kvist, Duneld, & Conway, 2013).

Ontologies are an outstanding contribution to biomedical informatics (Ivanović & Budimac, 2014). However, being verified and validated by a closed set of medical scientists imposes several drawbacks (Hoehndorf, Dumontier, Oellrich, Rebholz-Schuhmann, Schofield, & Gkoutos, 2011). On the one hand, an ontology can lack an important concept or relation that is found in other same-purpose biomedical ontologies (Hoehndorf, Dumontier, Oellrich, Rebholz-Schuhmann, Schofield, & Gkoutos, 2011). On the other hand, several existing ontologies are incompatible with one other because they cover different fields of interest or because they use different systems (Hoehndorf, Dumontier, Oellrich, Rebholz-Schuhmann, Schofield, & Gkoutos, 2011; Bodenreider, 2008). Moreover, existing databases have different licences to use (LTU). Many are subscription-based business models where additional technical barriers and affordability becomes a concern (Walsh, Arora, & Cohen, 2002). These issues limit integration of biomedical data and prevent semantic interoperability between biomedical computer systems (Bodenreider, 2008; Orgun & Wu, 2006). This has led to a fractured landscape of partial databases that are isolated from each other as well as the wider knowledge ecosystem, thus falling short of the potential of a broader knowledge-base. There is, therefore, a clear need for a single centralized, voluminous, easily editable and verifiable biomedical ontology, to which all biomedical concepts, relations, and all distinct biomedical ontologies should be compatible (Hoehndorf, Dumontier, Oellrich, Rebholz-Schuhmann, Schofield, & Gkoutos, 2011).

The innovative feature of Wikidata comes from its centralized and linked functionality and parametric approach. We are, in fact, involved in the research focusing on the use of Wiki resources (such as Wikipedia and Wiktionary) in relation to their semantics dimensions. Recently, we have been exploring the added value of Wikidata to the biomedical domain. As it stands, the biomedical databases are incomplete due to the lack of a cohesive platform that updates and records the ever-evolving daily changes and discoveries of scientific research. As new updates of scientific knowledge replace and modify older understandings, a single centralized update of the relevant Wikidata item can automatically update across multiple platforms and languages. It therefore has the potential to involve existing biomedical databases and integrate them with relevant information not yet contained in any of those existing ontologies. Here, we raise awareness of the potential use of Wikidata as a useful resource for biomedical data integration and semantic interoperability between biomedical computer systems (Jacobsen, Waagmeester, Kaliyaperumal, Stupp, Schriml, Thomspon, Su, & Roos, 2018).

Definitions

Large-scale database	A database including a voluminous set of data
Multilingual database	A database supporting data in multiple natural languages
Machine-readable database	A database that can be processed by computer methods and programs
Human-readable database	A database that can be read and understood by users
Structured database	A database in which data is always organized in the same format
Multidisciplinary database	Non-specialized database covering many fields of interest
Centralized database	A database that gets its data from other external databases and serves as a central reference for the others in terms of maintenance, logical structure, terms, etc.

Semantic database	A database that uses a defined language set and data structure in order to be more easily integrated with other resources
Knowledge-base	A database structured as a network of statements, rather data tables, to optimise storage and retrieval of complex information
Scholia	A tool (https://tools.wmflabs.org/scholia/) that creates profiles for publications, research areas, individuals, institutions, journals and publishers based on bibliographic information available in Wikidata (Nielsen, Mietchen, & Willighagen, 2017)

Wikidata as a large-scale semantic framework

In October 2012, Wikidata (www.wikidata.org) was created as an open, free, easily editable, ontological and collaborative semantic MediaWiki-based[2] knowledge-base to support Wikipedia (Vrandečić & Krötzsch, 2014; Vrandecic, 2013). It has subsequently grown in use and has become a large-scale, human-readable and machine-readable, multilingual, multidisciplinary, centralized, ontological knowledge-base (Vrandečić & Krötzsch, 2014). It currently supports over 350 languages and dialects[3] (Vrandečić & Krötzsch, 2014; Turki, Vrandečić, Hamdi, & Adel, 2017), particularly English, Dutch, French and German[4] (Kaffee, Piscopo, Vougiouklis, Simperl, Carr, & Pintscher, 2017), contains more than 200 million statements on about 56 million items and has a higher edit frequency than Wikipedia. Its rapid growth has been enabled by three main factors. Firstly, it has incorporated and combined several existing ontologies and datasets as well as the information already existing on Wikipedia. Secondly, it is easily modified and expanded thanks to its user and bot friendly interface with more than 18,000 active editors (Vrandečić & Krötzsch, 2014; Müller-Birn, Karran, Lehmann, & Luczak-Rösch, 2015). Thirdly, its Creative Commons CC0 license has facilitated its widespread use by other systems (Vrandečić & Krötzsch, 2014).

Wikidata has multiple required features for a scientific knowledge-base, such as the ability to provide references and qualifiers for single claims, the easy live update procedure, and the possibility to audit every edit through public registers (available at https://www.wikidata.org/wiki/Special:RecentChanges) (Vrandečić & Krötzsch, 2014). Wikidata connects to numerous external authorities, in a variety of areas (Vrandečić & Krötzsch, 2014). It also increasingly references research papers to support its statements, summarizing information from the literature into its machine-readable structure. Wikidata's architecture is based on widely used web standards, such as XML, JSON and RDF for representing its content (Vrandečić & Krötzsch, 2014), and SPARQL as a query language (Bielefeldt, Gonsior, & Krötzsch, 2018). Because of its regular structure, Wikidata can be easily used as the underlying database for other tools, such as Google Knowledge graph, or incorporated into other databases, such as DBpedia (Paulheim, 2017; Pellissier Tanon, Vrandečić, Schaffert, Steiner, & Pintscher, 2016; Ringler & Paulheim, 2017).

Data organization

As a large-scale knowledge graph, Wikidata consists of two types of entities: items (e.g.: https://www.wikidata.org/wiki/Q193216 for *Trachoma* [Q193216]) and properties (e.g.: https://www.wikidata.org/wiki/Property:P2176 for *drug used for treatment* [P2176]) (Vrandečić & Krötzsch, 2014). Items are the concepts in Wikidata ontology. As for properties, they are the types of relations between items. They can either be taxonomic relation types (e.g. *instance of* P31, *subclass*

[2]Further information about Semantic MediaWiki can be found in https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki.

[3]Labels, descriptions and aliases of entities are mainly added to Wikidata from online databases like Wikipedia and sister projects, available offline resources such as the lexicon of medicinal plants in Tunisian Arabic (Boukef, 1986) and Massive Open Online Courses (MOOC, E.g. https://tinyurl.com/y67n8qjg).

[4] https://www.wikidata.org/wiki/User:Pasleim/Language_statistics_for_items

of P279*, part of* P361), non-taxonomic relation types (e.g. *drug used for treatment* P2176), database matching relation types (e.g. P698 for *PubMed ID*), or ontology alignment relation types (e.g. *MeSH ID* P486) (Vrandečić & Krötzsch, 2014). Properties cannot be directly created by regular users to prevent duplication and disorganization of Wikidata properties. Instead, users willing to create a new property should make a proposal about it to the Wikidata community (e.g. Risk factor property proposal https://www.wikidata.org/wiki/Wikidata:Property_proposal/risk_factor). Every entity is assigned a label, a description and a list of aliases in multiple languages as well as a unique identifier.

This identifier is a number with prefix Q for items and a number with prefix P for properties (Vrandečić & Krötzsch, 2014). For each item, a list of claims is attributed in the form of triples (**Fig. 1**). The subject of the triple is the Wikidata item to which the claim refers, the predicate is a Wikidata property, and the object is a value. A value can be another item, a string, a URL, a time, a period, a location, or a quantity, depending on the property type. Claims can be made more precise through the use of qualifiers. These qualifiers show the contexts of the validity of the claim. Claims can be annotated through the inclusion of references. Qualifiers and references are also represented in the form of triples where the subject is the claim. To distinguish between qualifiers and references, the prefix P of the identifier of the Wikidata property of a reference triple is substituted by a prefix S. A claim and its references are considered a statement (Burgstaller-Muehlbacher, et al., 2016).

Wikidata is a secondary database, which means that it relies mainly on other resources to develop its content. It is not a data repository to add original work. To avoid errors done by new editors, each property has its own constraints related to its nature (single or multiple values, unit, mandatory qualifiers, etc.) (Hanika, Marx, & Stumme, 2019). For example, the property *strand orientation* P2548 for a gene on a double stranded DNA molecule accepts only two values: *forward strand* (Q22809680) and *reverse strand* (Q22809711). Those constraints are explained in the discussion page of the property (e.g. https://www.wikidata.org/wiki/Property_talk:P2548) and potential violations or vandalisms are reviewed by advanced users. When gathering data form different resources, a content harmonization should also be done (Landis, 2019). This issue has already emerged when implementing the Virtual International Authority File (VIAF) – an international reference file for authors and books that includes bibliographic and subject metadata (Loesch, 2011) – when different libraries had contradictory information (e.g. different data of birth of a painter) (Jin & Kudeki, 2019; Brenden Hansen, Hensten, Pedersen, & Bognerud, 2019). Further information about Wikidata data organization can be found at https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer.

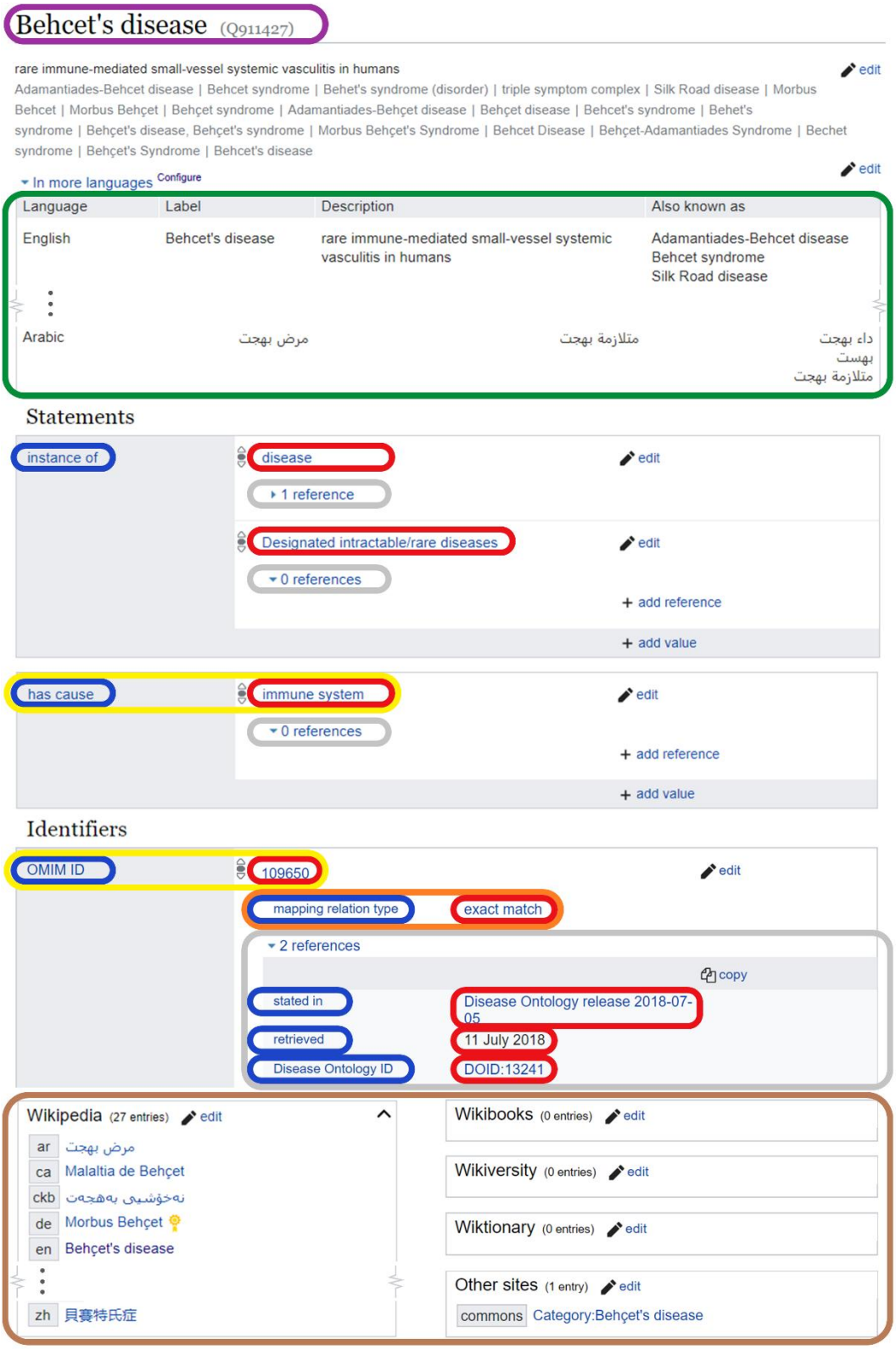


Figure 1: **Data structure of Wikidata.** A Wikidata item includes an identifier (purple), a list of labels, descriptions and aliases in multiple languages (green), sitelinks to Wikimedia projects related pages (brown) and a list of statements composed of claims (yellow), qualifiers (orange) and references (gray). Claims, references and qualifiers are triples where the predicate (blue) is a Wikidata property and the object (red) is a value, external ID, URL, date, string or another Wikidata item. Some items have been cropped in the image for clarity. This particular item currently has names in listed 42 languages, 62 Statements, 25 identifiers, and entries in 28 other Wikimedia projects as of May 23 2019 (Source: https://www.wikidata.org/wiki/Q911427, Note that exact layout will differ depending on screen width).

Automated data usage and enrichment

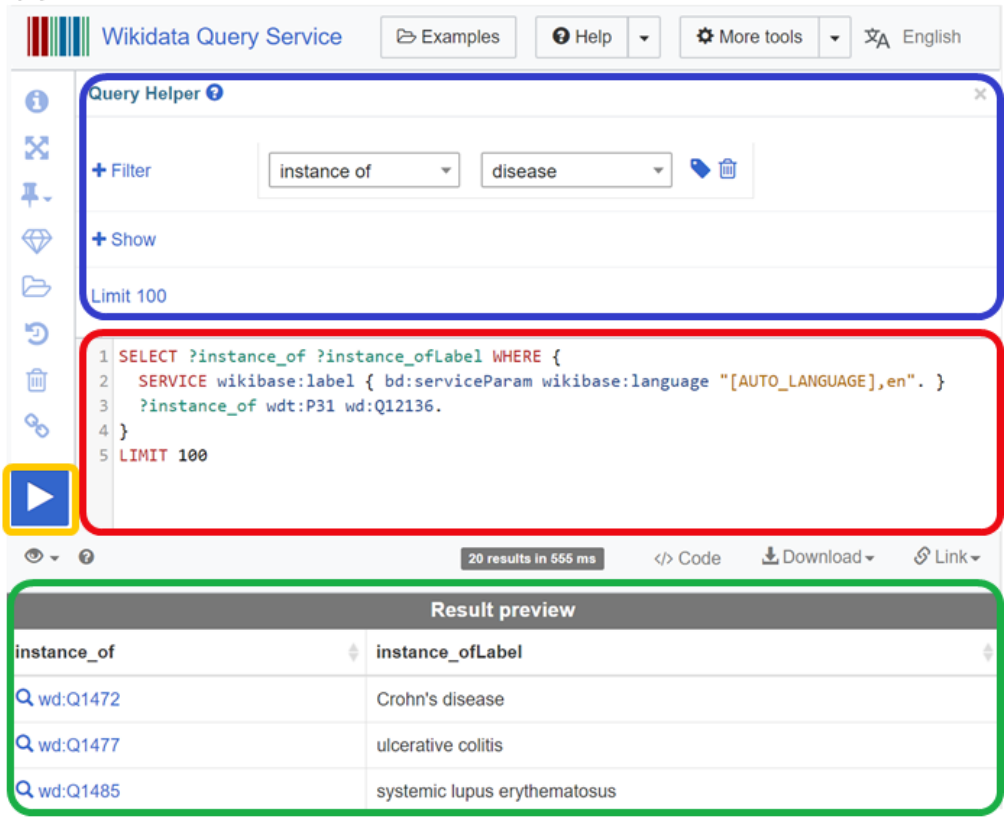
Further than the possibility to manually explore and edit Wikidata through its user-friendly interface (https://www.wikidata.org, Fig. 1), Wikidata can be automatically accessed and adjusted by computer methods and programs using a variety of methods (Burgstaller-Muehlbacher, et al., 2016).

The whole Wikidata knowledge-base can be easily downloaded in RDF, XML or JSON format using database dumps freely available at https://www.wikidata.org/wiki/Wikidata:Database_download (Färber, Ell, Menne, & Rettinger, 2015). The full record of a given Wikidata entity can be easily accessed in computer-friendly formats such RDF or JSON using the Linked Data Interface (Vrandečić & Krötzsch, 2014). To find the record in a computer-readable format <file_format> of a Wikidata entity having <entity_ID> as an identifier, it is only required to access https://www.wikidata.org/wiki/Special:EntityData/<entity_ID>.<file_format>. Applications can use the Wikidata API (Application programming interface) to find, query or edit data in Wikidata (Vrandečić & Krötzsch, 2014). This API and its detailed documentation are accessible at https://www.wikidata.org/w/api.php. Wikidata can be browsed or edited by bots, mainly written in Python using the Pywikibot framework (Pfundner, Schönberg, Horn, Boyce, & Samwald, 2015). Further information about Pywikibot can be found at https://www.mediawiki.org/wiki/Manual:Pywikibot. However, to run a bot on Wikidata, the user should register a separate Wikimedia account for the bot and create a request for permission to validate its automated edits. That request should include a full description of bot functions and its main purposes and architecture[5] (Pfundner, Schönberg, Horn, Boyce, & Samwald, 2015).

The interface is designed such that computer programming skills are useful but not necessary – since users can still automatically enrich Wikidata with their own dataset or knowledge or query it to answer questions, including medical ones (Mitraka, Waagmeester, Burgstaller-Muehlbacher, Schriml, Su, & Good, 2015). Effectively, the user can use the SPARQL Endpoint of Wikidata[6] (Mitraka, Waagmeester, Burgstaller-Muehlbacher, Schriml, Su, & Good, 2015). Queries can either be made by directly writing SPARQL query codes, or via a simple "query helper" graphical user interface (**Fig. 2A**). Clicking "Execute query" will then get query results. When the results are displayed, the user can visualize them as a table, a tree, a map, a chart or a graph but also save them in JSON, TSV, CSV, HTML and SVG formats. To get a list of which variants of which gene predict a positive prognosis in colorectal cancer, simple lines of SPARQL codes (available at https://w.wiki/4B8) are needed (**Fig. 2B, 2C**). By replacing Q188874 with Q128581, results will be shown for breast cancer instead (**Fig. 2D**). A key feature of Wikidata is a focus on connecting information from multiple fields. Such concepts could therefore be expanded to start asking far more complex queries such as "Which genes are associated with at least one type of cancer, and act in the same pathway as the targets of known drugs. Which researchers have published articles on these within the last 5 years?".

[5] E.g. https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/AlepfuBot
[6] https://query.wikidata.org

A



B

```
SELECT ?geneLabel ?variantLabel
WHERE
{
    VALUES ?disease {wd:Q188874}
    ?variant wdt:P3358 ?disease ; # P3358 Positive prognostic predictor
         wdt:P3433 ?gene . # P3433 biological variant of
    SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en" }
}
```

C

geneLabel	variantLabel
CDX2	CDX2 EXPRESSION
MIR218-1	MIR218-1 EXPRESSION
DCC	DCC EXPRESSION
POLE	POLE P286R
POLE	POLE V411L
POLE	POLE S459F
BRAF	BRAF Non-V600

D

geneLabel	variantLabel
PGR	PGR EXPRESSION
PIK3CA	PIK3CA MUTATION

Figure 2: **Interface of Wikidata SPARQL query service. A)** Example shown used to generate a list of diseases (https://query.wikidata.org): SPARQL query field (red), the query helper (blue), the execute query button (orange), the first three query results (green). This query currently returns 11685 results (as on May 23, 2019). Note that exact layout will differ depending on screen width. **B)** Query on which variants of which gene predict a positive prognosis in colorectal cancer. **C)** Query result from panel B, returning 7 results (as on May 23, 2019). **D)** Query result of which variants of which gene predict a positive prognosis in breast cancer, returning 2 results (as on May 23, 2019).

Similarly, there are several tools for the automatic enrichment of Wikidata without the need for coding (Nielsen, Mietchen, & Willighagen, 2017). The most important ones are QuickStatements and Descriptioner. QuickStatements[7] is used for creating items and/or adding statements to items and Descriptioner[8] is used to add new descriptions to all items that match a particular SPARQL query. For both of these, a Wikimedia account is required. To use QuickStatements, the user has to define each new statement or item in a single input line. A new statement should include an item (subject), a property (predicate), and a value (object) and can optionally involve qualifiers and references. After that, the user just clicks on "Do it" button to let the tool automatically insert the new statements or items to Wikidata. To use the Descriptioner tool, the user should simply enter a SPARQL query to find the Wikidata items to which a common description in a given natural language should be added. Then, the user should add the desired description and indicate its language before previewing and running the description process.

Wikidata as a valuable medical resource

When it comes to linking existing biomedical databases, major issues emerge. FAIR (Findable, Accessible, Interoperable, Reusable) data guidelines are challenging when there are different ontologies and controlled vocabularies due to technical and legal concerns. Even with similar databases (OBO, Bioschemas and SIO) results are inconclusive (Jacobsen, Waagmeester, Kaliyaperumal, Stupp, Schriml, Thomspon, Su, & Roos, 2018).

Just as Wikipedia has become a valuable resource of public health literacy (Heilman, et al., 2011; Shafee, Masukume, Kipersztok, Das, Häggström, & Heilman, 2017) and medical education (Azzam, et al., 2017), Wikidata is well-placed to take on the role as a centralized secondary biomedical ontological database (Vrandečić, 2012; Jacobsen, Waagmeester, Kaliyaperumal, Stupp, Schriml, Thomspon, Su, & Roos, 2018). In fact, it is taking a hub role in a growing number of fields and had achieved already significant outreach to the biology domain, in particular as a common and shared vocabulary for genome projects (Burgstaller-Muehlbacher, et al., 2016; Putnam, et al., 2017) or as a drug-drug interaction database (Ayvaz, et al., 2015). This provides an example of how versatile it can be as centralized knowledge-base to ensure biomedical data integration and interoperability between biomedical computer methods and programs. Furthermore, the breadth of Wikidata also enables biomedical internationalized resource identifiers (IRI) to be linked to the wider database, such as relationships between those items and pharmaceutical companies, researchers, historical events, or locations.

As it is a multidisciplinary large-scale ontological database, Wikidata involves many medicine-related entries although several types of medical items such as classifications[9], formulas[10], medical signs (e.g. back pain), valves, anatomical parts (e.g. Iliac crest), tests, radiological signs, EEG and ECG biomarkers (e.g. S1Q3 is an ECG biomarker for pulmonary embolism), diagnosis methods (e.g. Valsalva maneuver, Köplik's spots) and surgical procedures are still missing (**Fig. 3A**). These items range from human genes and proteins to anatomical entities such as arteries and veins and the most important ones have corresponding articles in the four largest language editions in Wikipedia (**Fig. 3B**). All these items are linked together to form an extended biomedical taxonomy using taxonomic Wikidata properties: *instance of* (P31), *subclass of* (P279), *part of* (P361) and *has part* (P527). This taxonomy is a key point as it allows computation of semantic similarity and relatedness in medical field (Pedersen, Pakhomov, Patwardhan, & Chute, 2007). This taxonomy enhances also the classification, annotation and hierarchical categorization of multilingual biomedical resources such as research papers (Vanteru, Shaik, & Yeasin, 2008; Xuan, et al., 2009), medical questions (Yu, Sable, & Zhu, 2005), electronic health records (Albright, et al., 2013), medical images (Seifert, et al., 2010) and DNA

[7] https://tools.wmflabs.org/wikidata-todo/quick_statements.php
[8] https://tools.wmflabs.org/pltools/descriptioner
[9] E.g. Gharbi ultrasonographic classification of hydatic cysts (Gharbi, Hassine, Brauner, & Dupuch, 1981).
[10] E.g. MDRD formula for estimating glomerular filtration rate (Levey, et al., 2009).

sequences (Osborne, et al., 2009). This will be useful for a variety of purposes like biomedical data mining (Koopman, Russell, & Zuccon, 2018; Xuan, et al., 2009) and the extraction of biomedical relations (Abulaish & Dey, 2007) and term definitions (Ma & Distel, 2013).

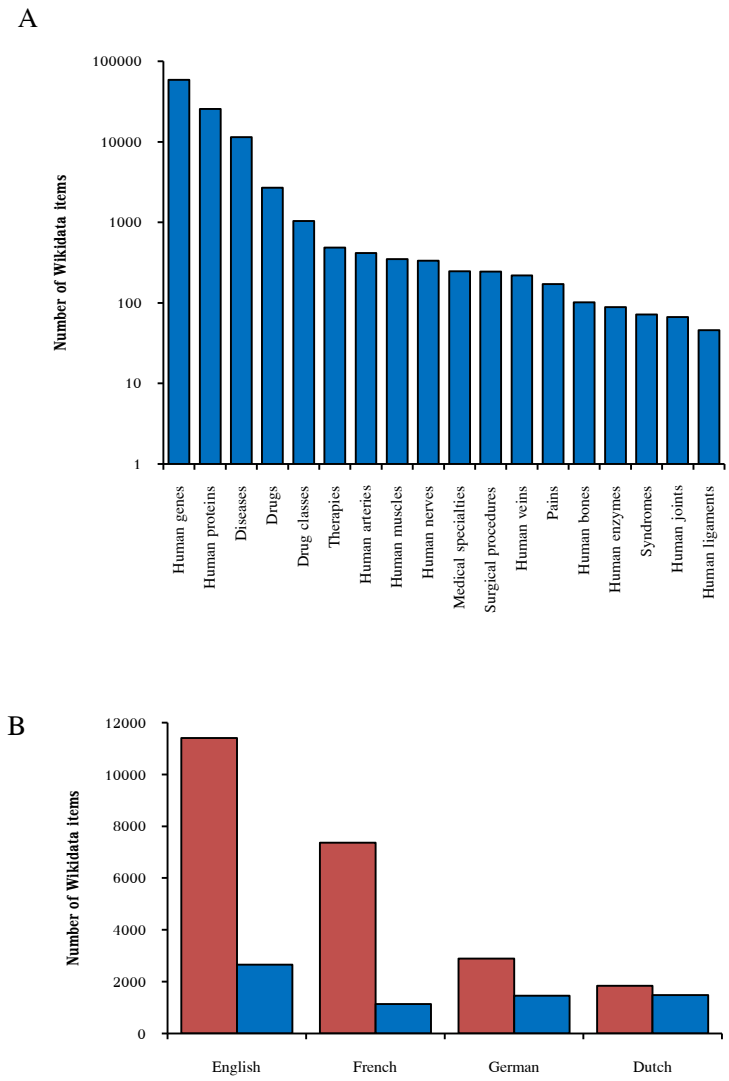


Figure 3: Medicine-related Wikidata items. A) Number of items by type. **B)** Number of items labelled as diseases (red) and drugs (blue) with articles in the four largest language editions of Wikipedia (As per https://meta.wikimedia.org/wiki/List_of_Wikipedias).

Wikidata does not only involve taxonomic relations but also non-taxonomic relations (Appendix A). Indeed, biomedical Wikidata items can be linked to other ones using biomedical relations such as *route of administration* (P636), *symptoms* (P780), and *side effect* (P1909). Such information is important as it is the structured representation of clinical knowledge and is consequently useful for clinical data processing (e.g. The interpretation of biological analyses and X-ray images) and decision support (e.g. The automation of medical diagnoses and of drug prescription and the identification of drug-drug interactions within therapies) (Bodenreider, 2008) and for biomedical question answering (**Fig. 4**) (Diefenbach, Tanon, Singh, & Maret, 2017).

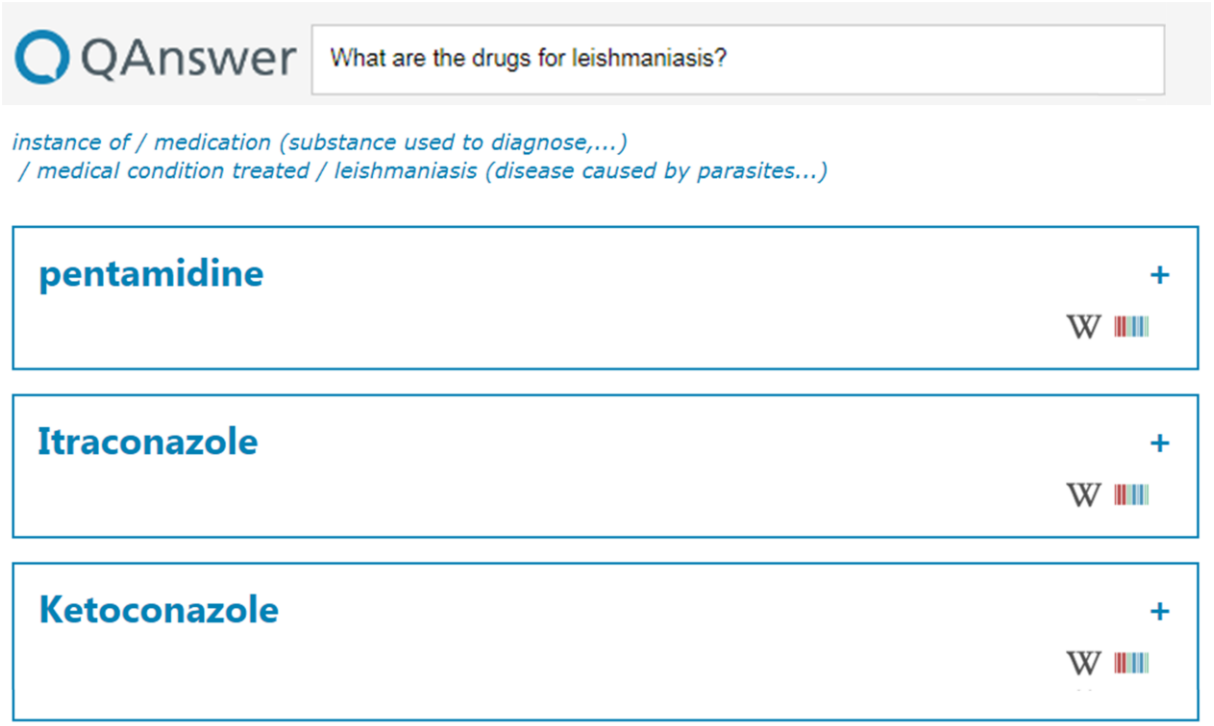


Figure 4 : Principle of Wikidata-based question answering (https://qanswer-frontend.univ-st-etienne.fr). Natural language questions are converted into machine-readable triples based on a dataset of sample questions. Then, SPARQL endpoint of Wikidata is used to return answers.

These medical taxonomic and non-taxonomic triples therefore act as a machine-readable version of human-readable biomedical sentences. For example, a pharmaceutical interaction can be described (**Example 1**). Similarly, Wikidata triplets can be used to describe elements of a schematic diagram and their supporting references (**Example 2**).

Example 1: Paracetamol is subclass of non-opioid analgesic that inhibits COX1 and is used to treat pain, but has a significant interaction with carbamazepine.

- Triple 1 = *paracetamol* (Q57055)/*subclass of* (P279)/*non-opioid analgesic* (Q1747785) (Cherny & Christakis, 2011)
- Triple 2 = *paracetamol* (Q57055)/*enzyme inhibitor* (Q427492)/*COX1* (Q21126810)[11]
- Triple 3 = *paracetamol* (Q57055)/*medical condition treated* (P2175)/*pain* (Q81938)[12]
- Triple 4 = *paracetamol* (Q57055)/*significant drug interaction* (P769)/*carbamazepine* (Q410412) (Cazacu, Mogosan, & Loghin, 2015)

[11]National Drug File, ID: N0000145898. http://bioportal.bioontology.org/ontologies/NDFRT?p=classes&conceptid=N0000145898

[12]IUPHAR/BPS Guide to pharmacology, ligand: 5239.http://www.guidetopharmacology.org/GRAC/LigandDisplayForward?ligandId=5239

Example 2: In the case of a eukaryotic gene, the 5' cap is a type of post transcriptional modification that is the part of a mature mRNA that binds ribosomes (Shafee & Lowe, 2017b).

- Triple 1A = *five prime cap* (Q238406)/*part of* (P361)/*gene* (Q7187) (Moore, 2005)
- Triple 1B = *gene* (Q7187)/of (P642)/*eukaryote* (Q19088) (Moore, 2005)
- Triple 2A = *five prime cap* (Q238406)/*part of* (P361)/*messenger RNA* (Q188928) (Guhaniyogi & Brewer, 2001)
- Triple 2B = *messenger RNA* (Q188928)/*of* (P642)/*eukaryote* (Q19088) (Guhaniyogi & Brewer, 2001)
- Triple 3 = *five prime cap* (Q238406)/*instance of* (P31)/*post-transcriptional modification* (Q417379) (Topisirovic, Svitkin, Sonenberg, & Shatkin, 2011)
- Triple 4A = *five prime cap* (Q238406)/*physically interacts with* (P129)/*ribosome* (Q42244) (Ramanathan, Robb, & Chan, 2016)
- Triple 4B = *ribosome* (Q42244)/*of* (P642)/*eukaryote* (Q19088) (Ramanathan, Robb, & Chan, 2016)

Each triple in the example sentences can be supported by one or more references, in a way that would be unwieldy if distributed through a human-readable sentence, and impractical when annotating parts of a diagram (Good, Burgstaller-Muehlbacher, Mitraka, Putman, Su, & Waagmeester, 2016). These statements then form an extended network that can be used to link concepts across medicine. Such networks can be visualised with the Wikidata Query Service. In the case of example 1, extending this to graph all *significant interactions* of drugs listed as being used to treat pain shows how extensive, yet still incomplete this dataset is (**Fig. 5A**). For example, the interaction between fentanyl and amiodarone was missing (now corrected online) (Pérez-Mañá, Papaseit, Fonseca, Farré, Torrens, & Farré, 2018). Similarly, extending the network of relationships around example 2 creates a network that can capture key relationships in the organisation of a gene's structure (**Fig. 5B**) in a machine-readable equivalent of pre-existing human-readable diagram representations (Shafee & Lowe, 2017b).

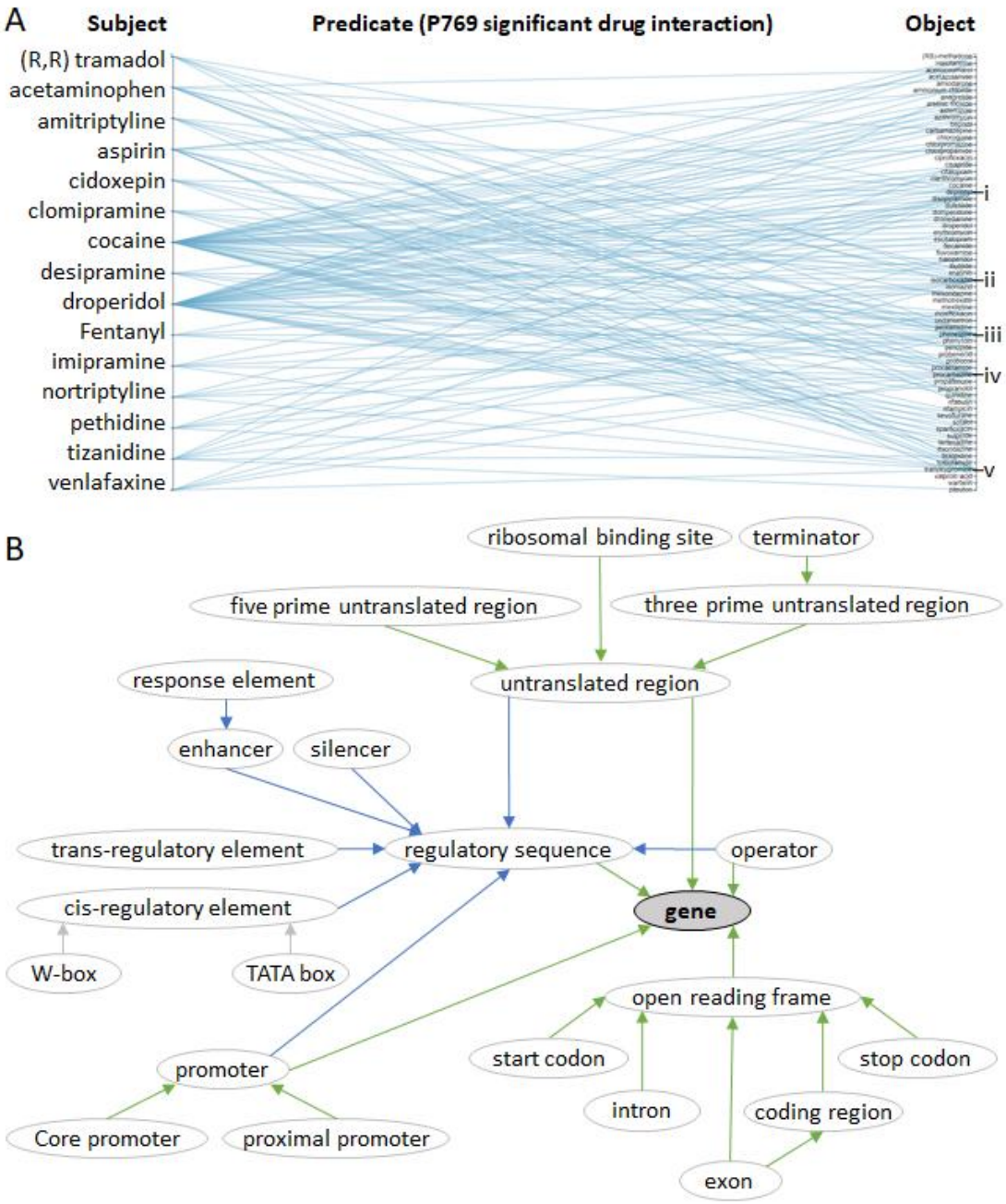


Figure 5. Examples of data visualisations using Wikidata query service. A) Network of all significant pharmacological interactions listed on Wikidata for drugs used in the treatment of pain. Edges indicate that the drug on the left axis (subject) has a significant drug interaction (P769) with the drug on the right axis (object). The most highly linked to are i) deprenyl, ii) isocarboxazid, iii) phenelzine, iv) procarbazine, and v) tranylcypromine. Labels on right hand axis in order: (RS)-methadone, halofantrine, acenocoumarol, acetazolamide, amiodarone, ammonium chloride, anagrelide, arsenic trioxide, astemizole, azithromycin, bepridil, carbamazepine, chloroquine, chlorpromazine, chlorpropamide, ciprofloxacin, cisapride, citalopram, clarithromycin, cocaine, deprenyl, disopyramide, dofetilide, domperidone, dronedarone, droperidol, erythromycin, escitalopram, flecainide, fluvoxamine, haloperidol, ibutilide, imatinib, isocarboxazid, isoniazid, mesoridazine, methotrexate, mexiletine, moxifloxacin, ondansetron, pentamidine, phenelzine, phenytoin, pimozide, probenecid, probucol, procainamide, procarbazine, propafenone, propranolol, quinidine, rifabutin, rifampicin, sevoflurane, sotalol, sparfloxacin, sulpiride, terfenadine, thioridazine, ticlopidine, tolbutamide, tranylcypromine, valproic acid, warfarin, zileuton. Live, interactive version at https://w.wiki/4Gh. **B) Edges** representing the predicates *part of (green)*, *subclass of (blue)* and *instance of* (grey) descending from the item for *Gene*. Live, interactive version at https://tinyurl.com/yatp7cdy

Wikidata also includes database identifiers like *Disease Ontology ID* (P699), *MeSH ID* (P486) and *SNOMED-CT identifier* (P5806) that link out biomedical items to the identifiers of equivalent items in external databases. Such relations ensure biomedical ontology alignment and interoperability between biomedical computer methods and programs based on different semantic resources. Moreover, Wikidata links biomedical items to their epidemiological statistics such as *incidence* (P2844), *prevalence* (P1193) and *disease burden* (P2854) allowing users to statistically analyse the health situation of nations and regions for health policy management purposes using one of the automated data usage tools explained before (Ramatowski, et al., 2017).

These features allowed Wikidata to cover quite all the subfields of medical science and to become comparable to major biomedical ontologies as shown in Table 1.

Table 1: Comparison of major biomedical language resources (Ivanović & Budimac, 2014; Kibbe, et al., 2014; Gene Ontology Consortium, 2016)

Name (Acronym)	Domain	Use	Relation types	Number of Elements	URL
Standardized Nomenclature of Medicine (SNOMED)	Clinical Medicine	EHR Documentation	Taxonomic relations	311000 concepts	http://www.snomed.org/
Medical Subject Heading (MeSH)	Biomedical indexing	Bibliographic Retrieval	Taxonomic relations	24767 concepts	https://www.ncbi.nlm.nih.gov/mesh
International Classification of Diseases (ICD-10)	Diseases	Public health support	Taxonomic relations	14000 classes	https://www.who.int/classifications/icd/en/
Unified Medical Language System (UMLS)	Interoperable biomedical information systems and services	Information Extraction	Taxonomic relations, Lexical relations	Over 1000000 concepts	https://www.nlm.nih.gov/research/umls/
Generalized Architecture for Languages, Encyclopedias and Nomenclatures in medicine (GALEN)	Anatomy and Clinical Medicine	EHR Documentation	Taxonomic relations	Over 25000 concepts	http://www.opengalen.org
Open Biomedical Ontologies (OBO)	All aspects of medicine, Interoperable biomedical information systems and services	Information Extraction and Processing	Taxonomic relations (relating entities from the same ontology), Non-taxonomic relations (relating entities from distinct ontologies)	60 specific-purpose ontologies each including thousands of concepts (e.g. Gene ontology)	http://www.obofoundry.org
Wikidata (WD)	All aspects of medicine	Information Extraction and Processing	Taxonomic and non-taxonomic relations	Over 100000 concepts. Also involves data about hospitals, species, physicians and medical tools	https://www.wikidata.org

Ensuring that all items and statements exist and are supported by correct references remains a large, ongoing task (Färber, Bartscherer, Menne, & Rettinger, 2018). In terms of quality, medicine-related items contain an average of 13 statements. These statements partly lack organization as shown in Fig. 6 (Brasileiro, Almeida, Carvalho, & Guizzardi, 2016). Of these statements, 88.11% are backed by references such as URL links or direct references to documents or databases (e.g. Fig. 7 for genetic association with AIDS). However, this fact does not reflect the real situation of the statements related to biomedical items in Wikidata. In fact, this important overall referencing rate is strongly influenced by the large number of gene and protein items, whose statements are >90% referenced thanks to Gene Wiki initiative (Burgstaller-Muehlbacher, et al., 2016; Putnam, et al., 2017). Existing statements related other types of biomedical items are in part not available (Except drugs and drug classes due to the work of Ayvaz et al. (Ayvaz, et al., 2015)) and not supported by references (Table 2). Given its increasing prominence, ensuring participation by clinicians, researchers and content

experts is increasingly valuable (Shafee, Mietchen, & Su, 2017c; Masukume, Kipersztok, Das, Shafee, Laurent, & Heilman, 2016).

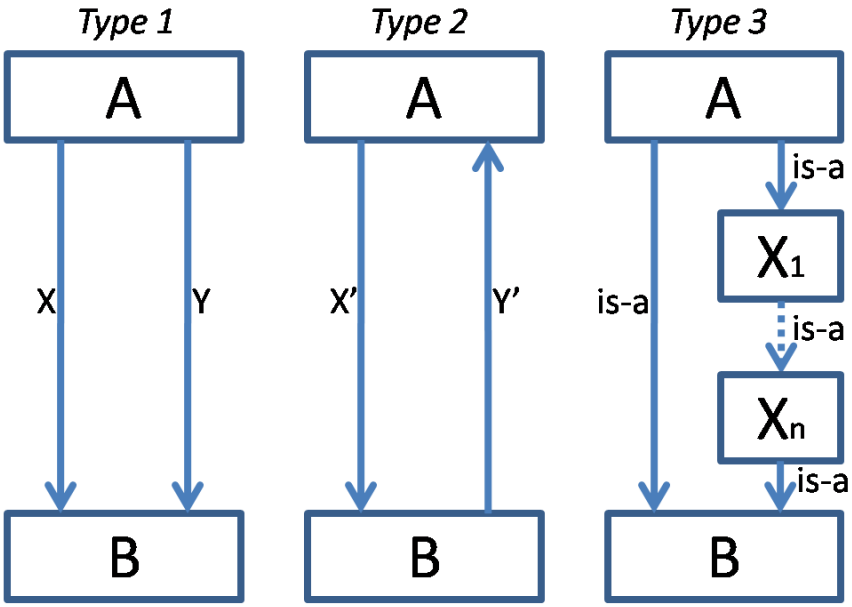


Figure 6: Common deficiencies related to the definition of biomedical Wikidata statements. A and B are biomedical items, X, Y, X' and Y' are unrelated Wikidata properties, and *is-a* is *instance of* (P31) or *subclass of* (P279). For examples, please see https://tinyurl.com/y48mxns4 (Type 1), https://tinyurl.com/y65jw89a (Type 2), and https://tinyurl.com/y2jd5eoc (Type 3).

Figure 7: The use of references to support Wikidata statements related to biomedical items. Wikidata reference for the genetic association of PARD3B with AIDS.

Table 2: Number of supported and unsupported statements for each type of biomedical entities (As on March 5, 2019, Best supported types are shown in bold)

Biomedical entity (P31)	Number of items	Number of properties		Number of properties per item		Percentage of referenced data
		With references	Without references	With references	Without references	
Drugs	2713	75259	35302	27.7	13.0	68.1%
Drug classes	1043	16855	10537	16.2	10.1	61.5%
Human enzymes	89	1234	386	13.9	4.3	76.2%
Diseases	11447	152622	57689	13.3	5.0	72.6%
Human genes	**58691**	**671282**	**12949**	**11.4**	**0.2**	**98.1%**
Human proteins	**25482**	**265684**	**27825**	**10.4**	**1.1**	**90.5%**
Human muscles	351	1690	2136	4.8	6.1	44.2%
Pains	171	725	858	4.2	5.0	45.8%
Syndromes	72	173	350	2.4	4.9	33.1%
Human arteries	418	964	2383	2.3	5.7	28.8%
Human joints	67	151	535	2.3	8.0	22.0%
Human bones	102	233	1119	2.3	11.0	17.2%
Human nerves	335	738	1738	2.2	5.2	29.8%
Human veins	220	478	1081	2.2	4.9	30.7%
Medical specialties	248	512	2069	2.1	8.3	19.8%
Therapies	487	931	2312	1.9	4.7	28.7%
Human ligaments	46	56	201	1.2	4.4	21.8%
Surgical procedures	244	261	1099	1.1	4.5	19.2%
Overall	102226	1189848	160569	11.6	1.6	88.1%

Conclusion and future work

It is clear that Wikidata is a valuable open data platform for biomedical data processing, integration and alignment. Due to its reading and editing flexibility, we believe it has the potential to take on a role as a centralized repository for medical knowledge as its accuracy is consequently fundamental to public health. We therefore advise a larger involvement from clinicians and researchers as well as biomedical institutions in Wikidata development. In terms of the future direction of this work, we plan to create tools to facilitate access and management of biomedical data within Wikidata and to compare it with other existing biomedical databases in order to assess the quality of the biomedical information provided by Wikidata. Furthermore, we will try to improve the organization of medical data within Wikidata by adding description logics support to Wikidata properties (Krötzsch, Simancik, & Horrocks, 2014) and taking into consideration the different names given by anatomists and surgeons to entities, the changes in the names of biomedical entities over the years (AlRyalat, et al., 2018), the morphological and physiological characteristics of anatomical entities, and the chronological evolution of the symptoms of diseases[13]. Moreover, we will upgrade Wikidata's biomedical data model so that it supports other types of useful clinical items, which will allow weight to be given to biomedical associations according a scale of importance[14]. We will further discuss the use of references of clinical statements in Wikidata to create structured and machine-readable abstracts of biomedical research publications[15] and the creation of an automatic process to enrich and adjust medical Wikidata using PubMed, MeSH and Scholia by identifying missing Wikidata properties related to medicine, by adding, verifying and adjusting biomedical items and taxonomic

[13] E.g. Acute appendicitis is characterized by an epigastric pain in early stages and by a right lower quadrant pain in late stages (Sherman, 1990).

[14] E.g. A pathognomonic sign of a disease should not have the same weight as a non-specific sign of that disease. As well, gold standards to cure or identify diseases should not be considered the same as other treatments or diagnosis methods.

[15] E.g. The *Supports the following statement(s)* section of https://tools.wmflabs.org/scholia/work/Q48672086.

and non-taxonomic statements and by adding missing references to Wikidata statements (Turki, Hadj Taieb, & Ben Aouicha, 2018) and verifying existing ones[16].

Acknowledgements

We thank Ms. Raquel Paredes for proofreading the work. We also thank the following for their useful comments and discussions that contributed to the preparation and refinement of this manuscript:

- Mr. Daniel Mietchen (University of Virginia, United States of America)
- Ms. Netha Hussain (University of Gothenburg, Sweden)
- Mr. James M. Heilman (University of British Columbia, Canada)
- Mr. Sujoy Ray (National Institute of Mental Health and Neurosciences, India)
- Mr. Finn Årup Nielsen (Technical University of Denmark, Denmark)
- Mr. Abel Lifaefi Mbula (Institut Supérieur des Techniques Médicales de Kisangani, DR Congo)
- Mr. Rajeeb Dutta (Wikimedia India, India)
- Ms. Maria Gessa (Universidad de Sevilla, Spain)
- Ms. Nancy Gertrudiz (Instituto Politécnico Nacional, Mexico)
- Mr. Vivek Podder (University of Dhaka, Bangladesh)
- Mr. Abhijeet Safai (Symbiosis International University, India)
- Mr. Alaa Najjar (University of Alexandria, Egypt)
- Mr. Varshil Mehta (Seven Hills Hospital, India)
- Mr. Mossab Banat (Hashemite University, Jordan)
- Mr. Sami Mlouhi (University of Monastir, Tunisia)
- Ms. Reem Al Kashif (Ain Shams University, Egypt)
- Mr. Scott MacLeod (World University and School, United States of America)
- Mr. Andy Mabbett (Wikimedia UK, United Kingdom)
- Mr. Kamel Boukef (University of Monastir, Tunisia)
- Ms. Mina Theofilatou (University of Patras, Greece)

Conflict of interest

All the authors of this research work are members of WikiProject Medicine, the group of individuals involved in the representation of medical knowledge in Wikidata.

References

Abulaish, M., & Dey, L. (2007). Biological relation extraction and query answering from MEDLINE abstracts using ontology-based text mining. *Data & Knowledge Engineering, 61* (2), 228-262.

Albright, D., Lanfranchi, A., Fredriksen, A., Styler IV, W. F., Warner, C., Hwang, J. D., et al. (2013). Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association, 20* (5), 922-930.

AlRyalat, S. A., Rawashdeh, K., Yasin, A., Alqadi, F., Saleh, N., Malkawi, L., et al. (2018). The change from an eponym to a representative name: Wegener to granulomatosis with polyangiitis. *Scientometrics, 117* (3), 2077-2089.

Athenikos, S. J., & Han, H. (2010). Biomedical question answering: A survey. *Computer methods and programs in biomedicine, 99* (1), 1-24.

[16] Recent clinical trials, systematic reviews and meta-analyses should be considered as better than other types of research publications in supporting biomedical Wikidata statements.

Ayvaz, S., Horn, J., Hassanzadeh, O., Zhu, Q., Stan, J., Tatonetti, N. P., et al. (2015). Toward a complete dataset of drug–drug interaction information from publicly available sources. *Journal of biomedical informatics, 55*, 206-217.

Azzam, A., Bresler, D., Leon, A., Maggio, L., Whitaker, E., Heilman, J., et al. (2017). Why medical schools should embrace Wikipedia: final-year medical student contributions to Wikipedia articles for academic credit at one School. *Academic Medicine, 92* (2), 194-200.

Bielefeldt, A., Gonsior, J., & Krötzsch, M. (2018). Practical Linked Data Access via SPARQL: The Case of Wikidata. *Proceedings of the WWW 2018 Workshop on Linked Data on the Web (LDOW-18).* Lyon: CEUR Workshop Proceedings.

Bodenreider, O. (2008). Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of medical informatics, 17* (1), 67-79.

Boukef, M. K. (1986). *Les plantes dans la médecine traditionnelle tunisienne.* Tunis: Agence de coopération culturelle et technique.

Brasileiro, F., Almeida, J. P. A., Carvalho, V. A., & Guizzardi, G. (2016, April). Applying a multi-level modeling theory to assess taxonomic hierarchies in Wikidata. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 975- 980). International World Wide Web Conferences Steering Committee.

Brenden Hansen, Y., Hensten, D., Pedersen, G. B., & Bognerud, M. (2019). Norwegian Artist Names Authority List of Artists in Norwegian Art Collections. *Heritage, 2* (1), 490-506.

Burgstaller-Muehlbacher, S., Waagmeester, A., Mitraka, E., Turner, J., Putman, T., Leong, J., et al. (2016). Wikidata as a semantic framework for the Gene Wiki initiative. *Database, 2016*, baw015.

Cazacu, I., Mogosan, C., &Loghin, F. (2015). Safety issues of current analgesics: an update. *Clujul Medical, 88* (2), 128-136.

Cherny, N. I., & Christakis, N. A. (2011). *Oxford textbook of palliative medicine – Chapter 7.* Oxford University Press.

Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D., Barrero, R. A., et al. (2006). A Multilingual Ontology for Infectious Disease Surveillance: Rationale, Design and Challenges. *Language Resources and Evaluation, 40* (3-4), 405-413.

De Potter, P., Cools, H., Depraetere, K., Mels, G., Debevere, P., De Roo, J., et al. (2012). Semantic patient information aggregation and medicinal decision support. *Computer Methods and Programs in Biomedicine, 108* (2), 724-735.

De Silva, T. S., MacDonald, D., Paterson, G., Sikdar, K. C., & Cochrane, B. (2011). Systematized nomenclature of medicine clinical terms (SNOMED CT) to represent computed tomography procedures. *Computer Methods and Programs in Biomedicine, 101* (3), 324-329.

Diefenbach, D., Tanon, T., Singh, K., &Maret, P. (2017). Question answering benchmarks for wikidata. *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks.* Vienna: CEUR Workshop Proceedings.

Färber, M., Bartscherer, F., Menne, C., & Rettinger, A. (2018). Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web, 9* (1), 77-129.

Färber, M., Ell, B., Menne, C., & Rettinger, A. (2015). A comparative survey of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web Journal, 1*, 1-5.

Gene Ontology Consortium. (2016). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic acids research, 45* (D1), D331-D338.

Gharbi, H. A., Hassine, W., Brauner, M. W., &Dupuch, K. (1981). Ultrasound examination of the hydatic live. *Radiology, 139* (2), 459-463.

Good, B. M., Burgstaller-Muehlbacher, S., Mitraka, E., Putman, T., Su, A. I., & Waagmeester, A. (2016). Opportunities and Challenges Presented by Wikidata in the Context of Biocuration. *International Conference on Biomedical Ontology and BioCreative (ICBO BioCreative 2016).* CEUR Workshop Proceedings.

Guhaniyogi, J., & Brewer, G. (2001). Regulation of mRNA stability in mammalian cells. *Gene, 265* (1-2), 11-23.

Hanika, T., Marx, M., & Stumme, G. (2019). Discovering Implicational Knowledge in Wikidata. *ArXiv preprint*, arXiv:1902.00916.

Heilman, J. M., Kemmann, E., Bonert, M., Chatterjee, A., Ragar, B., Beards, G. M., et al. (2011). Wikipedia: a key tool for global public health promotion. *Journal of medical Internet research, 13* (1), e14.

Henriksson, A., Skeppstedt, M., Kvist, M., Duneld, M., & Conway, M. (2013). Corpus-driven terminology development: populating Swedish SNOMED CT with synonyms extracted from electronic health records. *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, (pp. 36-44).

Hoehndorf, R., Dumontier, M., Oellrich, A., Rebholz-Schuhmann, D., Schofield, P. N., & Gkoutos, G. V. (2011). Interoperability between Biomedical Ontologies through Relation Expansion, Upper-Level Ontologies and Automatic Reasoning. *PLoS One, 6* (7), e22006.

Ivanović, M., & Budimac, Z. (2014). An overview of ontologies and data resources in medical domains. *Expert systems with applications, 41*, 5158-5166.

Jacobsen, A., Waagmeester, A., Kaliyaperumal, R., Stupp, G. S., Schriml, L. M., Thompson, M., Su, A. I., & Roos, M. (2018). Wikidata as an intuitive resource towards semantic data modeling in data FAIRification. *Proceedings of the 11th International Conference Semantic Web Applications and Tools for Life Sciences (SWAT4HCLS 2018)*. Antwerp: SWAT4LS.

Jin, Q., & Kudeki, D. (2019). Identity and Access Management for Libraries. *Technical Services Quarterly, 36* (1), 44-60.

Kaffee, L. A., Piscopo, A., Vougiouklis, P., Simperl, E., Carr, L., & Pintscher, L. (2017). A glimpse into Babel: an analysis of multilinguality in Wikidata. *Proceedings of the 13th International Symposium on Open Collaboration* (p. 14). ACM.

Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., et al. (2014). Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research, 43* (D1), D1071-D1078.

Koopman, B., Russell, J., &Zuccon, G. (2018). Task-oriented search for evidence-based medicine. *International Journal on Digital Libraries, 19* (2-3), 217-229.

Krötzsch, M., Simancik, F., & Horrocks, I. (2014). Description logics. *IEEE Intelligent Systems, 29*(1), 12-19.

Landis, C. (2019). Linked open data in libraries. In K. J. Varnum, *New top technologies every librarian needs to know* (pp. 3-15). Chicago: ALA Neal-Schuman.

Levey, A. S., Stevens, L. A., Schmid, C. H., Zhang, Y. L., Castro, A. F., Feldman, H. I., et al. (2009). A new equation to estimate glomerular filtration rate. *Annals of internal medicine, 150* (9), 604-612.

Liu, F., Fontelo, P., & Ackerman, M. (2006). BabelMeSH: development of a cross-language tool for MEDLINE/PubMed. *AMIA Annual Symposium Proceedings* (p. 1012). American Medical Informatics Association.

Loesch, M. F. (2011). VIAF (The Virtual International Authority File)–http://viaf.org. *Technical Services Quarterly, 28* (2), 255-256.

Ma, Y., & Distel, F. (2013). Learning formal definitions for SNOMED CT from text. *Conference on Artificial Intelligence in Medicine in Europe* (pp. 73-77). Berlin, Heidelberg: Springer.

Masukume, G., Kipersztok, L., Das, D., Shafee, T. M., Laurent, M. R., & Heilman, J. M. (2016). Medical journals and Wikipedia: a global health matter. *The Lancet Global Health, 4* (11), e791.

Mitraka, E., Waagmeester, A., Burgstaller-Muehlbacher, S., Schriml, L. M., Su, A. I., & Good, B. M. (2015). Wikidata: A platform for data integration and dissemination for the life sciences and beyond. *bioRxiv*, 031971.

Moore, M. J. (2005). From birth to death: the complex lives of eukaryotic mRNAs. *Science, 309* (5740), 1514-1518.

Müller-Birn, C., Karran, B., Lehmann, J., & Luczak-Rösch, M. (2015). Peer-production system or collaborative ontology engineering effort: What is Wikidata? *Proceedings of the 11th International Symposium on Open Collaboration* (p. 20). ACM.

Nielsen, F. Å., Mietchen, D., & Willighagen, E. (2017). Scholia, scientometrics and wikidata. *European Semantic Web Conference* (pp. 237-259). Springer, Cham.

Orgun, B., & Wu, J. (2006). HL7 ontology and mobile agents for interoperability in heterogeneous medical information systems. *Computers in Biology and Medicine, 36* (7-8), 817-836.

Osborne, J. D., Flatow, J., Holko, M., Lin, S. M., Kibbe, W. A., Zhu, L. J., et al. (2009). Annotating the human genome with Disease Ontology. *BMC genomics, 10* (Suppl 1), S6.

Patel, V. L., Shortiffe, E. H., Stefanelli, M., Szolovits, P., Berthold, M. R., et al. (2009). The coming of age of artificial intelligence in medicine. *Artificial intelligence in medicine, 46* (1), 5-17.

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web, 8* (3), 489-508.

Pedersen, T., Pakhomov, S. V., Patwardhan, S., & Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics, 40* (3), 288-299.

Pellissier Tanon, T., Vrandečić, D., Schaffert, S., Steiner, T., & Pintscher, L. (2016). From freebase to wikidata: The great migration. *Proceedings of the 25th international conference on world wide web* (pp. 1419-1428). International World Wide Web Conferences Steering Committee.

Pérez-Mañá, C., Papaseit, E., Fonseca, F., Farré, A., Torrens, M., & Farré, M. (2018). Drug interactions with new synthetic opioids. *Frontiers in pharmacology, 9*, 1145.

Pfundner, A., Schönberg, T., Horn, J., Boyce, R. D., & Samwald, M. (2015). Utilizing the Wikidata System to Improve the Quality of Medical Content in Wikipedia in Diverse Languages: A Pilot Study. *Journal of Medical Internet Research, 17* (5), e110.

Putnam, T. E., Lelong, S., Burgstaller-Muehlbacher, S., Waagmeester, A., Diesh, C., Dunn, N., et al. (2017). WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata. *Database, 2017*, bax025.

Ramanathan, A., Robb, G. B., & Chan, S.-H. (2016). mRNA capping: biological functions and applications. *Nucleic acids research, 44* (16), 7511-7526.

Ramatowski, J. W., Lee, C. X., Mantzavino, A., Ribas, J., Guerra, W., Preston, N. D., et al. (2017). Planning an innovation marathon at an infectious disease conference with results from the International Meeting of Emerging Diseases and Surveillance 2016 Hackathon. *International Journal of Infectious Diseases, 65*, 93-97.

Ringler, D., &Paulheim, H. (2017). One knowledge graph to rule them all? Analyzing the differences between DBpedia, YAGO, Wikidata & co. *Joint German/Austrian Conference on Artificial Intelligence (KünstlicheIntelligenz)* (pp. 366-372). Springer, Cham.

Seifert, S., Kelm, M., Moeller, M., Mukherjee, S., Cavallaro, A., Huber, M., et al. (2010). Semantic annotation of medical images. *Medical Imaging 2010: Advanced PACS-based Imaging Informatics and Therapeutic Applications.7628*, p. 762808. International Society for Optics and Photonics.

Shafee, T., & Lowe, R. (2017b). Eukaryotic and prokaryotic gene structure. *WikiJournal of Medicine, 4* (1), 2.

Shafee, T., Masukume, G., Kipersztok, L., Das, D., Häggström, M., & Heilman, J. (2017). Evolution of Wikipedia's medical content: past, present and future. *Journal of Epidemiology and Community Health, 71* (11), 1122-1129.

Shafee, T., Mietchen, D., &Su, A. I. (2017c). Academics can help shape Wikipedia. *Science, 357* (6351), 557-558.

Sherman, R. (1990). Abdominal Pain. In H. K. Walker, W. D. Hall, & J. W. Hurst, *Clinical Methods: The History, Physical, and Laboratory Examinations.* Boston: Butterworths.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., et al. (2005). Relations in biomedical ontologies. *Genome biology, 6* (5), R46.

Sun, P. G. (2015). The human drug-disease-gene network. *Information Sciences, 306*, 70-80.

Topisirovic, I., Svitkin, Y. V., Sonenberg, N., & Shatkin, A. J. (2011). Cap and cap-binding proteins in the control of gene expression. *Wiley Interdisciplinary Reviews: RNA, 2* (2), 277-298.

Turki, H., Vrandečić, D., Hamdi, H., & Adel, I. (2017). Using WikiData as a multi-lingual multi-dialectal dictionary for Arabic dialects. *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications* (pp. 437-442). Hammamet: IEEE.

Turki, H., Hadj Taieb, M. A., & Ben Aouicha, M. (2018). MeSH qualifiers, publication types and relation occurrence frequency are also useful for a better sentence-level extraction of biomedical relations. *Journal of Biomedical Informatics, 83*, 217-218.

Vanteru, B. C., Shaik, J. S., &Yeasin, M. (2008). Semantically linking and browsing PubMed abstracts with gene ontology. *BMC Genomics, 9* (Suppl 1), S10.

Vrandecic, D. (2013). The rise of Wikidata. *IEEE Intelligent Systems, 28* (4), 90-95.

Vrandečić, D. (2012). Wikidata: a new platform for collaborative data collection. *Proceedings of the 21st International Conference on World Wide Web* (pp. 1063-1064). Lyon: ACM.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM, 57* (10), 78-85.

Walsh, J. P., Arora, A., & Cohen, W. M. (2003). Research Tool Patenting and Licensing and Biomedical Innovation. In *Patents in the Knowledge-Based Economy* (pp. 285-340). National Academies Press.

Xu, R., Li, L., & Wang, Q. (2014). dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text. *BMC Bioinformatics, 15* (1), 105.

Xuan, W., Dai, M., Mirel, B., Song, J., Athey, B., Watson, S. J., et al. (2009). Open biomedical ontology-based Medline exploration. *BMC bioinformatics, 10* (5), S6.

Yu, H., Sable, C., & Zhu, H. R. (2005). Classifying medical questions based on an evidence taxonomy. *Proceedings of the AAAI 2005 workshop on question answering in restricted domains* (pp. 27-35). AAAI Press.

Appendix A: Exhaustive list of non-taxonomic properties used for biomedical items in Wikidata

ID	Property name	Description	Target data type
P636	route of administration	path by which a drug, fluid, poison, or other substance is taken into the body	Wikidata item
P688	encodes	the product of a gene (protein or RNA)	Wikidata item
P689	afflicts	type of organism which a condition or disease afflicts	Wikidata item
P702	encoded by	the gene that encodes some gene product	Wikidata item
P769	significant drug interaction	clinically significant interaction between two pharmacologically active substances (i.e., drugs and/or active metabolites) where concomitant intake can lead to altered effectiveness or adverse drug events.	Wikidata item
P780	symptoms	possible symptoms of a medical condition	Wikidata item
P923	medical examinations	examinations that might be used to diagnose the medical condition	Wikidata item
P924	medical treatment	treatment that might be used to heal the medical condition	Wikidata item
P925	presynaptic connection	neuron connects on its presynaptic end to	Wikidata item
P926	postsynaptic connection	neuron connects on its postsynaptic end to	Wikidata item
P927	anatomical location	where in the body does this anatomical feature lie	Wikidata item
P928	activating neurotransmitter	which neurotransmitter activates the neuron	Wikidata item
P970	neurological function	the function of a neurological structure	Wikidata item
P1050	medical condition	any state relevant to the health of an organism, including diseases and positive conditions	Wikidata item
P1057	chromosome	chromosome on which an entity is localized	Wikidata item
P1060	pathogen transmission process	process by which a pathogen is transmitted	Wikidata item
P1199	mode of inheritance	manner in which a particular genetic trait or disorder is passed from one generation to the next	Wikidata item
P1604	biosafety level	level of the biocontainment precautions required to isolate dangerous biological agents	Wikidata item
P1605	has natural reservoir	host species for the pathogen in which it is endemic	Wikidata item
P1606	natural reservoir of	pathogen of which this species is a long-term host	Wikidata item
P1660	has index case	initial patient in the population of an epidemiological investigation	Wikidata item
P1677	index case of	primary case, patient zero: initial patient in the population of an epidemiological investigation	Wikidata item
P1909	side effect	effect of a medication or procedure, that occurs next to the desired effect	Wikidata item
P1910	decreased expression in	This property should link a gene and a disease and indicate that a decreased expression of the gene is found in the disease	Wikidata item
P1911	increased expression in	this property should link a gene and a disease and indicate that an increased expression of the gene is found in the disease	Wikidata item
P1912	deletion association with	This property should link a gene and a disease due to a deletion	Wikidata item
P1913	gene duplication association with	This property should link a gene and a disease due to a duplication	Wikidata item
P1914	gene insertion association with	This property should link a gene and a disease due to an insertion	Wikidata item
P1915	gene inversion association with	This property should link a gene and a disease due to an inversion	Wikidata item
P1916	gene substitution association with	This property should link a gene and a disease due to a substitution	Wikidata item
P1917	posttranslational modification association with	This property should link a gene and a disease due to an altered posttranslational modification	Wikidata item
P1924	vaccine for	disease that a vaccine is for	Wikidata item

P1995	health specialty	main specialty that diagnoses, prevent human illness, injury and other physical and mental impairments	Wikidata item
P2175	medical condition treated	disease that this pharmaceutical drug, procedure, or therapy is used to treat	Wikidata item
P2176	drug used for treatment	drug, procedure, or therapy that can be used to treat a medical condition	Wikidata item
P2239	first aid measures	actions to take to help a person in the case of accidents, injuries and accidental exposures to harmful chemicals	Wikidata item
P2286	arterial supply	arterial supply of an anatomical structure	Wikidata item
P2288	lymphatic drainage	lymphatic drainage of an anatomical structure	Wikidata item
P2289	venous drainage	vein draining the anatomical structure	Wikidata item
P2329	antagonist muscle	Muscle performing the opposite action than the given muscle	Wikidata item
P2789	connects with	item with which the item is physically connected	Wikidata item
P2841	age of onset	age group in which disease manifestations appear	Wikidata item
P3094	develops from	this class of items develops from another class of items	Wikidata item
P3189	innervated by	nerves which innervate this anatomical structure	Wikidata item
P3190	innervates	anatomical structures innervated by this nerve	Wikidata item
P3205	patient of	was treated or studied as a patient by this person	Wikidata item
P3261	anatomical branch of	main stem of this blood vessel, lymphatic vessel or nerve	Wikidata item
P3262	has anatomical branch	branches of this blood vessel, lymphatic vessel or nerve	Wikidata item
P3310	muscle action	action a muscle engages in	Wikidata item
P3354	positive therapeutic predictor	the presence of the genetic variant helps to predict response to a treatment	Wikidata item
P3355	negative therapeutic predictor	the presence of the genetic variant helps to predict no response or resistance to a treatment	Wikidata item
P3356	positive diagnostic predictor	the presence of the genetic variant helps to diagnose the presence of disease, used as inclusion criteria	Wikidata item
P3357	negative diagnostic predictor	the presence of the genetic variant helps to diagnose the absence of disease, used as exclusion criteria	Wikidata item
P3358	positive prognostic predictor	the presence of the genetic variant helps to prognose good outcome for the disease	Wikidata item
P3359	negative prognostic predictor	the presence of the genetic variant helps to prognose poor outcome for the disease	Wikidata item
P3464	medicine marketing authorization	medicinal product's marketing authorization status	Wikidata item
P3489	pregnancy category	official categorisation of safety of medicine in pregnancy	Wikidata item
P3490	muscle origin	the anatomic entity to which the beginning of a muscle is anchored	Wikidata item
P3491	muscle insertion	the anatomic entity to which the end of a muscle is anchored	Wikidata item
P3493	legal status (medicine)	legal status for pharmaceutical drugs, e.g. general sales list for paracetamol in the UK	Wikidata item
P4044	therapeutic area	disease area in which a medical intervention is applied	Wikidata item
P4425	mtDNA haplogroup	mitochondrial DNA haplogroup of a person or organism	Wikidata item
P4426	Y-DNA Haplogroup	Y-DNA haplogroup of a person or organism	Wikidata item
P4545	sexually homologous with	body part that originates from the same tissue or cell during fetal development in the opposing sex	Wikidata item
P4777	has boundary	element that's on the two-dimensional border that surrounds the subject	Wikidata item
P4843	development of anatomical structure	biological development of this anatomical structure	Wikidata item
P4954	may prevent	disease which may be prevented by this substance	Wikidata item
P5131	possible medical findings	possible medical findings of a medical condition	Wikidata item
P5132	suggests the existence of	medical conditions associated with the subject medical finding	Wikidata item
P5133	has evaluation	evaluation or interpretation corresponding to the subject attribute or examination	Wikidata item
P5134	evaluation of	the subject finding is an evaluation or interpretation of the object attribute or test	Wikidata item
P5248	medical evacuation to	site to which those injured are evacuated immediately after a catastrophic event or a battle	Wikidata item
P5446	reference value	value or range of values for a medical test (subject item), used to evaluate whether the results of the test are normal or not	Wikidata item

P5572	expressed in	gene or protein is expressed during a specific condition/cell cycle/process/form	Wikidata item
P5642	risk factor	factor associated with a high prevalence of a medical condition	Wikidata item
P2293	Genetic association	general link between a disease and the causal genetic entity, if the detailed mechanism is unknown/unavailable	Wikidata item
P1193	prevalence	portion of a population with a given disease	Quantity
P1603	number of cases	cumulative number of confirmed, probable and suspected occurrences	Quantity
P2710	minimal lethal concentration	lowest concentration of a toxic substance in an environmental medium that kills individual organisms or test species under a defined set of conditions	Quantity
P2712	median lethal concentration	statistically derived median concentration of a substance in an environmental medium expected to kill 50% of organisms in a given population under a defined set of conditions	Quantity
P2717	no-observed-adverse-effect level	greatest concentration or amount of a substance, found by experiment or observation, which causes no detectable adverse alteration of morphology, functional capacity, growth, development, or life span of the target organism under defined conditions of exposure	Quantity
P2718	lowest-observed-adverse-effect level	lowest concentration or amount of a substance (dose), found by experiment or observation, which causes an adverse effect on morphology, functional capacity, growth, development, or life span of a target organism distinguishable from normal (control) organisms of the same species and strain under defined conditions of exposure	Quantity
P2844	incidence	probability of occurrence of a given condition in a population within a specified period of time	Quantity
P2854	disease burden	impact of a health problem as measured by financial cost, mortality, morbidity, or other indicators. It is often quantified in terms of quality-adjusted life years (QALYs) or disability-adjusted life years (DALYs)	Quantity
P3457	case fatality rate	proportion of patients who die of a particular medical condition out of all who have this condition within a given time frame	Quantity
P3487	maximal incubation period in humans	maximal time between an infection and the onset of disease symptoms in infected humans	Quantity
P3488	minimal incubation period in humans	minimal time between an infection and the onset of disease symptoms in infected humans	Quantity
P3492	basic reproduction number	number of infections caused by one infection within an uninfected population	Quantity
P4250	defined daily dose	average maintenance dose per day for a medicine used for its main indication in adults	Quantity
P2275	World Health Organisation International Nonproprietary Name	identifier for a drug	Monolingual text
P593	HomoloGene ID	identifier in the HomoloGene database	External ID
P667	ICPC 2 ID	classification method for primary care encounters	External ID
P1402	Foundational Model of Anatomy ID	identifier for human anatomical terminology	External ID
P1461	Patientplus ID	identifier of disease at Patient UK	External ID
P1692	ICD-9-CM	identifier in the ICD adaption assigning diagnostic and procedure codes	External ID
P1748	NCI Thesaurus ID	identifier in the United States National Cancer Institute Thesaurus, vocabulary for clinical care, translational and basic research, etc	External ID
P486	MeSH ID	identifier for concepts in the Medical Subject Headings controlled vocabulary	External ID
P492	OMIM ID	Online "Mendelian Inheritance in Man" catalogue codes for diseases, genes, or phenotypes	External ID
P493	ICD-9	identifier in the ICD catalogue codes for diseases - Version 9	External ID
P494	ICD-10	identifier in the ICD catalogue codes for diseases - Version 10	External ID

P557	DiseasesDB	identifier sourced on the Diseases Database	External ID
P563	ICD-O	International Classification of Diseases for Oncology	External ID
P592	ChEMBL ID	identifier from a chemical database of bioactive molecules with drug-like properties	External ID
P594	Ensembl Gene ID	identifier for a gene as per the Ensembl (European Bioinformatics Institute and the Wellcome Trust Sanger Institute) database	External ID
P595	Guide to Pharmacology Ligand ID	ligand identifier of the Guide to Pharmacology database	External ID
P604	MedlinePlus ID	health information from U.S. government agencies, and health-related organizations	External ID
P637	RefSeq Protein ID	identifier for a protein	External ID
P638	PDB structure ID	identifier for 3D structural data as per the PDB (Protein Data Bank) database	External ID
P639	RefSeq RNA ID	RNA Identifier	External ID
P652	UNII	identifier issued by the FDA / Unique Ingredient Identifier	External ID
P653	PubMed Health	identifier for a physiological condition, in the PubMed Health website	External ID
P663	DSM-IV	classification found in the Diagnostic and Statistical Manual of Mental Disorders	External ID
P665	KEGG ID	identifier from databases dealing with genomes, enzymatic pathways, and biological chemicals	External ID
P668	GeneReviews ID	collection of peer-reviewed articles that describe specific gene-related diseases	External ID
P672	MeSH Code	Medical Subject Headings (MeSH) codes are an index and thesaurus for the life sciences (≠ MeSH ID)	External ID
P673	eMedicine	online clinical medical knowledge base	External ID
P696	Neurolex ID	identifier in the Neurolex database	External ID
P698	PubMed ID	identifier for journal articles/abstracts in PubMed	External ID
P699	Disease Ontology ID	identifier in the Disease Ontology database	External ID
P704	Ensembl Transcript ID	transcript ID issued by Ensembl database	External ID
P715	Drugbank ID	identifier in the bioinformatics and cheminformatics database from the University of Alberta	External ID
P1055	NLM Unique ID	identifier in the catalog of the National Library of Medicine	External ID
P1323	TerminologiaAnatomica 98 ID	TerminologiaAnatomica (1998 edition) human anatomical terminology identifier	External ID
P1395	National Cancer Institute ID	identifier at www.cancer.gov	External ID
P1550	Orphanet ID	identifier in the Orpha.net database (without ORPHA prefix)	External ID
P1554	UBERON ID	identifier from UBERON ontology (without prefix)	External ID
P1583	MalaCards ID	identifier in the Malacards database of diseases	External ID
P1690	ICD-10-PCS	medical classification used for procedural coding	External ID
P1691	operations and procedures key (OPS)	official classification of operational procedures	External ID
P1693	TerminologiaEmbryologica	standardized list of words used in the description of human embryologic and fetal structures	External ID
P1694	TerminologiaHistologica	controlled vocabulary for use in cytology and histology	External ID
P1925	VIOLIN ID	identifier in the VIOLIN database for vaccines and related subjects	External ID
P1928	Vaccine Ontology ID	identifier in the Vaccine Ontology database	External ID
P1929	ClinVar Variation ID	identifier in the ClinVar database for human genomic variation	External ID
P1930	DSM-5	identifier for a mental disorder in the 5th edition of "Diagnostic and Statistical Manual of Mental Disorders"	External ID
P2074	internetmedicin.se ID	ID in the Swedish database about medical topics	External ID
P2646	mirTarBase ID	identifier for the mirTarBase database, a database for microRNAs and their targets	External ID
P2892	UMLS CUI	NLM Unified Medical Language System (UMLS) controlled biomedical vocabulary unique identifier	External ID
P2941	Munk's Roll ID	identifier of a person, in the biographical website of the Royal College of Physicians, London	External ID
P2944	Plarr ID	identifier of a person, in the biographical website of the Royal	External ID

		College of Surgeons, London	
P3098	ClinicalTrials.gov Identifier	identifier in the ClinicalTrials.gov database	External ID
P3201	Medical Dictionary for Regulatory Activities ID	identifier in the Medical Dictionary for Regulatory Activities	External ID
P3291	DocCheckFlexikonEn ID	Identifier for an article in the English section of DocCheckFlexikon wiki	External ID
P3292	DocCheckFlexikon De ID	Identifier for an article in the German section of the DocCheck wiki	External ID
P3329	CIViC variant ID	Identifier used in the CIViC database to identify specific variant	External ID
P3331	HGVS nomenclature	Sequence Variant Nomenclature from the Human Genome Variation Society (HGVS)	External ID
P3345	RxNorm CUI	Identifier for the normalized clinical drug dictionary of the Unified Medical Language System	External ID
P3550	Australian Register of Therapeutic Goods ID	identifier of a medicine or medical device listed in the Australian Register of Therapeutic Goods	External ID
P3637	European Medicines Agency product number	identifier issued by the European Medicines Agency for treatments approved in the European Union	External ID
P3640	National Drug Code	A pharmaceutical code issued by the Food and Drug Administration for every drug product (formulation) on the U.S. market. Includes a labeler code, product code and package code, unique for every drug product.	External ID
P3720	GPnotebook ID	ID of a topic, in the British medical database GPnotebook	External ID
P3841	Human Phenotype Ontology ID	The Human Phenotype Ontology (HPO) is a widely used vocabulary of phenotypic abnormalities encountered in human disease	External ID
P3885	History of Modern Biomedicine ID	identifier of a person or topic in the History of Modern Biomedicine database	External ID
P3945	RANM member ID	identifier of a member of the Spanish Royal Academy of Medicine	External ID
P3956	National Academy of Medicine (France) Member ID	identifier of a member of the French National Academy of Medicine	External ID
P3982	TA98 Latin term	Latin name for anatomical subject as described in TerminologiaAnatomica 98	External ID
P4058	FINESS medical facility ID	identifier of a medical facility in France in the FINESS directory	External ID
P4229	ICD-10-CM	classification of medical condition under ICD-10-CM (a country specific expansion of ICD-10)	External ID
P4233	PatientLikeMe condition ID	identifier for a condition on PatientsLikeMe, a website where patients can share health information	External ID
P4235	PatientLikeMe treatment ID	identifier for a treatment on PatientsLikeMe, a website where patients can share health information	External ID
P4236	PatientLikeMe symptom ID	identifier for a symptom on PatientsLikeMe, a website where patients can share health information	External ID
P4317	GARD rare disease ID	identifier for a rare disease in the United States National Institutes of Health's Genetic and Rare Diseases (GARD) Information Center database	External ID
P4338	LOINC ID	identifier for medical observations, measurements, and documents in the Regenstrief Institute's Logical Observation Identifiers Names and Codes (LOINC), a database of internationalized medical terminology	External ID
P4394	NeuroNames ID (plain mode)	numeric identifier of a brain structure in the BrainInfo (NeuroNames) database in plain mode	External ID
P4395	BrainInfo ID (hierarchical)	numeric identifier of a brain structure in the BrainInfo (NeuroNames) database in hierarchical mode	External ID
P4495	Xenopus Anatomical Ontology ID	identifier for an anatomical structure in the Xenopus Anatomical Ontology, a controlled vocubulary for describing Xenopus anatomy and embryological development	External ID
P4670	Sjukvårdsrådgivningen Category ID	identifier for a disease or medical condition, in the Swedish government's 'Sjukvårdsrådgivningen' database	External ID
P5082	Store medisinskeleksikon ID	identifier of an article in the online encyclopedia sml.snl.no	External ID

P5209	ISO 3950 code	identifier for a tooth per ISO 3950	External ID
P5270	MonDO ID	identifier for a disease in the Monarch Disease Ontology	External ID
P5329	ARMB member ID	Royal Academy of Medicine of Belgium member ID	External ID
P5375	BIU Santé person ID	identifier for a person on the Bibliothèque interuniversitaire de santé website	External ID
P5376	Medicina author ID	identifier for an author in the Medicina database	External ID
P5415	Whonamedit? doctor ID	identifier for a doctor on the Whonamedit? website	External ID
P5450	BIA PSY person ID	identifier for a person on the BiographischesArchiv der Psychiatrie website	External ID
P5458	Guide to Pharmacology Target ID	target identifier of the Guide to Pharmacology database	External ID
P5468	Historia de la Medicinaperson ID	identifier for a person on the Historia de la Medicina website	External ID
P5496	Médicoshistoricos doctor ID	identifier for a doctor in the Médicoshistoricos database of the Complutense University of Madrid	External ID
P5501	Brenda Tissue Ontology ID	identifier in Brenda Tissue Ontology for an enzyme source	External ID
P5806	SNOMED CT identifier	identifier in the SNOMED CT catalogue codes for diseases, symptoms and procedures	External ID
P5843	Gynopedia place ID	identifier for a place (city, region or country) on the Gynopedia wiki about sexual and reproductive health care	External ID
P6220	OpenTrials ID	identifier for OpenTrials database over clinical trials	External ID