Semantic Web -1 (2021) 1–32
DOI 10.3233/SW-210444
IOS Press

Representing COVID-19 information in collaborative knowledge graphs: The case of Wikidata

Houcemeddine Turki [a,b], Mohamed Ali Hadj Taieb [b], Thomas Shafee [c], Tiago Lubiana [d], Dariusz Jemielniak [e], Mohamed Ben Aouicha [f], Jose Emilio Labra Gayo [g], Eric A. Youngstrom [h], Mus'ab Banat [i], Diptanshu Das [j,k], Daniel Mietchen [l,m,*] and on behalf of WikiProject COVID-19 [**]

[a] *Faculty of Medicine of Sfax, University of Sfax, Sfax, Tunisia*
ORCID: https://orcid.org/0000-0003-3492-2014
[b] *Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia*
ORCID: https://orcid.org/0000-0002-2786-8913
[c] *La Trobe University, Melbourne, Victoria, Australia*
ORCID: https://orcid.org/0000-0002-2298-7593
[d] *Computational Systems Biology Laboratory, University of São Paulo, São Paulo, Brazil*
ORCID: https://orcid.org/0000-0003-2473-2313
[e] *Department of Management in Networked and Digital Societies, Kozminski University, Warsaw, Poland*
ORCID: https://orcid.org/0000-0002-3745-7931
[f] *Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia*
ORCID: https://orcid.org/0000-0002-2277-5814
[g] *Web Semantics Oviedo (WESO) Research Group, University of Oviedo, Spain*
ORCID: https://orcid.org/0000-0001-8907-5348
[h] *Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, CB #3270, Davie Hall, Chapel Hill, NC 27599-3270, United States of America*
ORCID: https://orcid.org/0000-0003-2251-6860
[i] *Faculty of Medicine, Hashemite University, Zarqa, Jordan*
ORCID: https://orcid.org/0000-0001-9132-3849
[j] *Institute of Child Health (ICH), Kolkata, India*
[k] *Medical Superspecialty Hospital, Kolkata, India*
ORCID: https://orcid.org/0000-0002-7221-5022
[l] *School of Data Science, University of Virginia, Charlottesville, Virginia, United States of America*
[m] *Biomedical Data & Bioethics, Fraunhofer Institute for Biomedical Engineering, Würzburg, Germany*
ORCID: https://orcid.org/0000-0001-9488-1870

CORRECTED PROOF

Editor: Armin Haller, Australian National University, Australia
Solicited reviews: Pouya Ghiasnezhad Omran, Australian National University, Australia; Gengchen Mai, Stanford University, USA; two anonymous reviewers

[*] Corresponding author. E-mail: daniel.mietchen@ibmt.fraunhofer.de.
[**] Project members: Jan Ainali, Susanna Ånäs, Erica Azzellini, Mus'ab Banat, Mohamed Ben Aouicha, Alessandra Boccone, Jane Darnell, Diptanshu Das, Lena Denis, Rich Farmbrough, Daniel Fernández-Álvarez, Konrad Foerstner, Jose Emilio Labra Gayo, Mauricio V. Genta, Mohamed Ali Hadj Taieb, James Hare, Alejandro González Hevia, David Hicks, Toby Hudson, Netha Hussain, Jinoy Tom Jacob, Dariusz Jemielniak, Krupal Kasyap, Will Kent, Samuel Klein, Jasper J. Koehorst, Martina Kutmon, Antoine Logean, Tiago Lubiana, Andy Mabbett, Kimberli Mäkäräinen, Tania Maio, Bodhisattwa Mandal, Nandhini Meenakshi, Daniel Mietchen, Nandana Mihindukulasooriya, Mahir Morshed, Peter Murray-Rust, Minh Nguyễn, Finn Årup Nielsen, Mike Nolan, Shay Nowick, Julian Leonardo Paez, João Alexandre Peschanski, Alexander Pico, Lane Rasberry, Mairelys Lemus-Rojas, Diego Saez-Trumper, Magnus Sälgö, John Samuel, Peter J. Schaap, Jodi Schneider, Thomas Shafee, Nick Sheppard, Adam Shorland, Ranjith Siji, Michal Josef Špaček, Ralf Stephan, Andrew I. Su, Hilary Thorsen, Houcemeddine Turki, Lisa M. Verhagen, Denny Vrandečić, Andra Waagmeester, and Egon Willighagen.

1570-0844 © 2021 – The authors. Published by IOS Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0).

Abstract. Information related to the COVID-19 pandemic ranges from biological to bibliographic, from geographical to genetic and beyond. The structure of the raw data is highly complex, so converting it to meaningful insight requires data curation, integration, extraction and visualization, the global crowdsourcing of which provides both additional challenges and opportunities. Wikidata is an interdisciplinary, multilingual, open collaborative knowledge base of more than 90 million entities connected by well over a billion relationships. It acts as a web-scale platform for broader computer-supported cooperative work and linked open data, since it can be written to and queried in multiple ways in near real time by specialists, automated tools and the public. The main query language, SPARQL, is a semantic language used to retrieve and process information from databases saved in Resource Description Framework (RDF) format.

Here, we introduce four aspects of Wikidata that enable it to serve as a knowledge base for general information on the COVID-19 pandemic: its flexible data model, its multilingual features, its alignment to multiple external databases, and its multidisciplinary organization. The rich knowledge graph created for COVID-19 in Wikidata can be visualized, explored, and analyzed for purposes like decision support as well as educational and scholarly research.

Keywords: Public health surveillance, Wikidata, knowledge graph, COVID-19, SPARQL, community curation, FAIR data, linked open data

CORRECTED PROOF

1. Introduction

The COVID-19 pandemic is complex and multifaceted and touches on almost every aspect of current life [25]. Coordinating efforts to systematize and formalize knowledge about COVID-19 in a computable form is key in accelerating our response to the pathogen and future epidemics [24]. There are already attempts at creating community-based ontologies of COVID-19 knowledge and data [37], as well as efforts to aggregate expert data [73]. Many open data initiatives have been started spontaneously [22,62,106]. The interconnected, multidisciplinary, and international nature of the pandemic creates both challenges and opportunities for using knowledge graphs [8,22–24,35,37,46,73,108]. However, there have been no systematic studies of crowd-sourced knowledge graph generation by spontaneous groups of self-coordinated users, under the pressure of rapidly occurring phenomena, such as the pandemic. Our paper fills this gap.

For applications of knowledge graphs in general, common challenges include the timely assessment of the relevance and quality of any piece of information with regards to the characteristics of the graph and the integration with other pieces of information within or external to the knowledge graph. Common opportunities are mainly related to leveraging such knowledge graphs for real-life applications, which in the case of COVID-19 could be, for instance, outbreak management in a specific societal context or education about the virus or about countermeasures [8,22–24,35,37,46,73,108]. While this manuscript as a whole emphasizes the opportunities, we think it is worthwhile to highlight some of the challenges early on.

1.1. COVID-19 data challenges

The integration of different data sources always poses a range of challenges [19], for example in terms of interoperability (e.g. differing criteria for COVID-19 deaths across jurisdictions), granularity (e.g. number of tests performed per jurisdiction and time period), quality control (e.g. whether aggregations of sub-national data fit with national data), data accessibility (e.g. whether they are programmatically and publicly accessible, and under what license) or scalability (e.g. how many sources to integrate, or how often to sync between them).

Integrating COVID-19 data presents particular challenges: First, human knowledge about the COVID-19 disease, the underlying pathogen and the resulting pandemic is evolving rapidly [53], so systems representing it need to be flexible and scalable in terms of their data models and workflows, yet quick in terms of deployability and updatability. Second, COVID-19-related knowledge, while very limited at the start of the pandemic, was still embedded in a broader set of knowledge (e.g. about viruses, viral infections, past disease outbreaks and interventions), and these relationships – which knowledge bases are meant to leverage – are growing along with the expansion of our

COVID-19 knowledge [105]. Third, the COVID-19 pandemic has affected almost every aspect of our globalized human society, so knowledge bases capturing information about it need to reflect that. Fourth, despite the disruptions that the pandemic has brought to many communities and infrastructures [25], the curated data about it should ideally be easily and reliably accessible for humans and machines across a broad range of use cases [82].

1.2. Organization of the manuscript

In this research paper, we report on the efforts of the Wikidata community (including our own) to meet the COVID-19 data challenges outlined in the previous section by using Wikidata as a platform for collaboratively collecting, curating and visualizing COVID-19-related knowledge at scales commensurate with the pandemic. While the relative merits of Wikidata with respect to other knowledge graphs have been discussed previously [1,30,84], we focus on leveraging the potential of Wikidata as an existing platform with an existing community in a timely fashion for an emerging transdisciplinary application like the COVID-19 response.

As active editors of Wikidata, the authors have contributed a significant part of that data modelling, usage framework and crowdsourcing of the COVID-19 information in the knowledge graph since the beginning of the pandemic. We consequently have a unique perspective to share our experience and overview how to use Wikidata to host COVID-19 data, integrate it with non-COVID-19 information and feed computer applications in an open and transparent way.

The remainder of the paper is organized as follows: we start by introducing Wikidata in general (Section 2) and describe key aspects of its data model in the context of the COVID-19 pandemic (Section 2.1). Then, we give an overview of the language support (Section 2.2) and database alignment (Section 2.3) of COVID-19 information in Wikidata. Subsequently, we present snapshots of applications of the Wikidata's COVID-19 knowledge graph to visualizing multidisciplinary information about COVID-19 (Section 3). These visualizations cover biological and clinical aspects (Section 3.1), epidemiology (Section 3.2), research outputs (Section 3.3) and societal aspects (Section 3.4). Finally, we discuss the outcomes of the open development of the COVID-19 knowledge graph in Wikidata (Section 4), draw conclusions and highlight potential directions for future research (Section 5).

2. Wikidata as a semantic resource for COVID-19

Wikidata is a large-scale, collaborative, open-licensed, multilingual knowledge base that is both human- and machine-readable. Notably, it is available in the standardized RDF (Resource Description Framework) format, where data is organized into entities (items) and the relationships that connect them to each other and outside data, named properties [102].

Wikidata is a peer production project, developed under the umbrella of the Wikimedia Foundation, which also hosts Wikipedia and an ecosystem of open collaborative websites around it. Similarly to Wikipedia, it relies on community-driven development and design and is both a-hierarchical and largely uncoordinated [47]. As a result, it develops entirely organically, based on the editor community's consensus, which may be implicit (e.g. by the absence of modifications) or explicit (e.g. a policy on how to handle biographical information about living people). This community develops ontologies and typologies used in the database.

This community-centric approach is both a blessing and a curse. On the one hand, it makes methodical planning of the whole structure and its granularity very difficult, if not impossible [59]: there simply is no central coordination system, and all major design decisions have to be approved through a consensus of all interested contributors. On the other hand, harnessing knowledge and skills of a broad range of human and automated contributors provides for an unparalleled flexibility and versatility of uses, and allows for rapid addressing of emerging and urgent phenomena, such as disease outbreaks.[1]

The novelty of a bottom-up developed Knowledge Graph relies on an entirely organic growth of taxonomies and content, negotiated continuously by the involved parties. While the benefits of peer production and collaborative editing are well known, they are particularly visible in contemporary and fast changing topics [54]. This is because

[1] https://www.wikidata.org/wiki/Wikidata:WikiProject_Humanitarian_Wikidata.

CORRECTED PROOF

the crowd-sourced coordination does not require a long decision-making process, nor a chain of command. Additionally, the bottom-up approach allows for a better optimization of topics, by relying on "free market" spontaneous forces of individual editors. It is already known that the search habits of users seeking medical content changed dramatically as a result of the COVID pandemic [13]. However, the exact dynamics of how this peer network responded to the challenge, in particular to the urgent need for new taxonomies and knowledge graphs, has not been a topic of systematic analysis. Our paper fills this gap.

With respect to the COVID-19 data challenges (cf. Section 1.1), Wikidata addresses them in several ways: First, it was designed for web scale data with flexible and evolving data models that can be updated quickly and frequently [98,102], and its existing community has been using it to capture COVID-19-related knowledge right from the start.[2] Second, Wikidata already contained a considerable and continuously expanding volume of curated background information – from SARS-CoV-1 and other coronaviruses to zoonoses, cruise ships, public health interventions, vaccine development and relevant publications – ready to be leveraged to explore the growing COVID-19-related knowledge in such broader contexts [105]. Third, both the Wikidata platform and the Wikidata community are highly multifaceted, multilingual and multidisciplinary [51,52]. Fourth, the Wikidata infrastructure is digital-first, with high uptime and low access barriers, while its community is distributed around the globe and includes people from many walks of life [102], such that any particular disruption due to the pandemic only affects subsets of the Wikidata community, which also has experience with handling humanitarian crises, e.g. through the Zika experience [28] and through overlap with the Wikipedia community that has been covering disasters for two decades.[3]

An important caveat is that data integration through Wikidata poses some particular challenges of its own, such as data licensing (being in the public domain, Wikidata can essentially only ingest public-domain data [76]) or multilinguality (e.g. how to handle concepts that are hard to translate [88]), and for certain kinds of data (e.g. health data from individual patients), it is not suitable, although appropriately configured instances of the underlying technology stack might [80].

Here, we present how various types of data related to the COVID-19 pandemic are currently represented in Wikidata thanks to the flexible structure of the database and how useful visualizations for different subsets of the data linked to COVID-19 within the Wikidata knowledge base can be generated.

2.1. *Data model flexibility*

In Wikidata, each concept has an item (a human, disease, drug, city, etc.) that is assigned a unique identifier (Q-number; brown in Fig. 1), and optionally a label, description and aliases in multiple languages (yellow in Fig. 1). The assignment of a single language-independent identifier for each entity in Wikidata helps minimize the size of the knowledge graph and avoids issues seen in databases such as DBpedia, where separate items are needed for each language [1]. Such a feature is allowed thanks to the use of Wikibase software – a MediaWiki variant adapted to support structured data – to drive Wikidata instead of other systems that represent entities using textual expressions, particularly Virtuoso in the context of DBpedia [1] and NewsReader [101].

The true richness of the knowledge base comes from the connections between the items: statements in the form of RDF triples (subject-predicate-object) where the subject is the respective item, the predicate is a Wikidata property (red in Fig. 1), and the object is another Wikidata item or piece of information (blue in Fig. 1). The properties that relate items are similarly each assigned an identifier (P-number). Some properties relate a Wikidata item as the object and can be taxonomic (e.g. *instance of* [P31], *subclass of* [P279] or *part of* [P361]) or non-taxonomic (e.g. *significant person* [P3342], *drug used for treatment* [P2176] or *symptoms* [P780]). Conversely, other properties can

[2]The creation dates of the three core items: "COVID-19 pandemic" (Q81068910) 2020-01-05 https://w.wiki/3PCe (note that labels may change), "SARS-CoV-2" (Q82069695) 2020-01-14 https://w.wiki/3PCf and "COVID-19" (Q84263196) 2020-02-02 https://w.wiki/3PCi. The creation and curation of these core items was accompanied by the creation of more specialized ones like "Category:COVID-19 pandemic" (Q83189805; 2020-01-19), "Wuhan Huanan Seafood Wholesale Market" (Q83264280; 2020-01-20), "Timeline of the COVID-19 pandemic" (Q83493517; 2020-01-24) or "Huoshenshan Hospital" (Q83554783; 2020-01-24).

As these items became available, they were quickly put to use for enriching the knowledge graph around them. For instance, when the paper "Recent advances in the detection of respiratory virus infection in humans" was published on 2020-01-15, the item Q82838328 about it had been linked to the "SARS-CoV-2" item within less than three days: https://w.wiki/3XAt.

[3]Cf. https://w.wiki/VDe.

CORRECTED PROOF

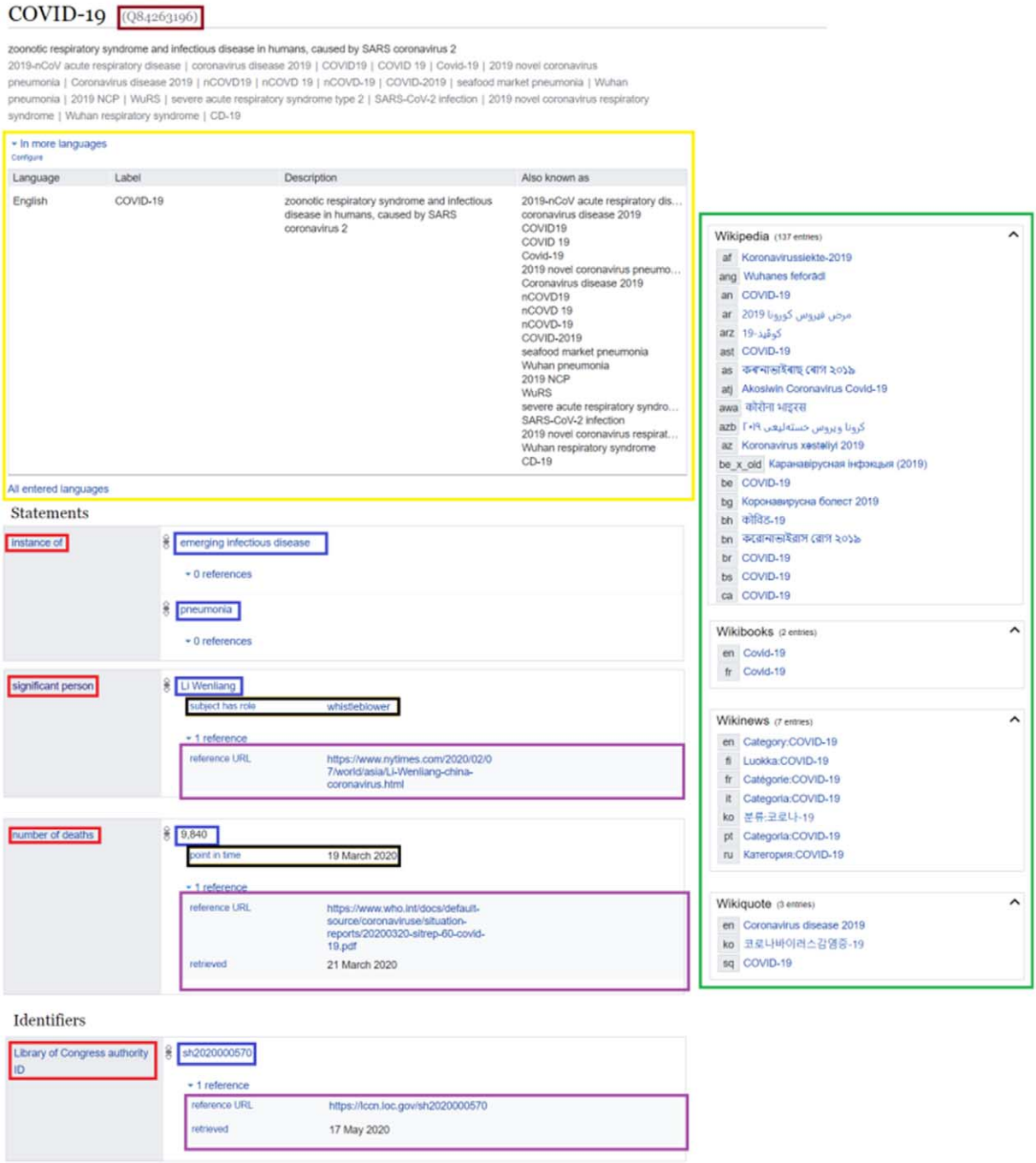


Fig. 1. Data structure of a Wikidata item. The simple, consistent structure of a Wikidata item makes it both human- and machine-readable. Each Wikidata item has a unique identifier (brown). Items can have labels, descriptions and aliases in multiple languages (yellow). They can include any number of statements having predicates (red), objects (blue), qualifiers (black) and references (purple) where the subject is the item. Finally, where additional wikimedia resources are available about an item's topic, those are listed (green). Source: https://www.wikidata.org/wiki/Q84263196, available at: https://w.wiki/auF. License: CC BY-SA 4.0.

have an object that is a value (e.g. *number of cases* [P1603]), date (e.g. *point in time* [P585]), URL (e.g. *official website* [P856]), string (e.g. *official name* [P1448]), or external identifier (e.g. *Library of Congress authority ID* [P244] or *Disease Ontology ID* [P699]). Each statement can be given further detail and specificity via qualifiers (black in Fig. 1) or provenance via references (purple in Fig. 1), which themselves can be represented as RDF triples [98]. This process of adding statements about statements is generally known as reification [40] and differs from other projects like DBpedia, which are based on a simple RDF representation without qualifiers and references [45]. The lack of reification in DBpedia makes it difficult to analyze epidemiological information as time series because values are not assigned dates. On the other hand, the excessive use of generic properties, particularly dbo:wikiPageWikiLink, to link DBpedia items according to how their corresponding Wikipedia articles are related in Wikipedia (e.g. wikilink or subcategory) does not always capture semantic relatedness as generic relations can sometimes represent comparisons and simultaneity and does not seem to be useful to drive automatic biomedical reasoning requiring knowledge resources with well-defined relation types to efficiently work. An example of such a deficiency is the consideration in DBpedia of a link to COVID-19 pandemic in Tunisia from the Wikipedia article about 2020 Nice Stabbing although the statement including the wikilink declares that the perpetrator had arrived as a migrant in late September 2020 at the island of Lampedusa, Italy, amidst the COVID-19 pandemic in Tunisia.

The only situation where DBpedia retrieves precise relational statements (e.g. dbp:symptoms, dbp:treatment) as well as non-relational statements (e.g. dbo:confirmedCases, dbp:arrivalDate) from Wikipedia is when the information is extracted from infoboxes [83]. Even in this situation, the infobox-based creation of DBpedia statements suffers from several inconsistencies requiring the use of logical constraints and human efforts for their efficient elimination although properties are quite defined in infoboxes [83]. A practical example of this problem is the DBpedia item about Ahmed Al-Qadri, a former Syrian minister of Agriculture and Agrarian Reform, as of June 20, 2021.[4] This item is described as having COVID-19 pandemic as a place of death (dbp:deathPlace) and not as a cause of death.

The advantage of Wikidata's use of RDF over other competing semantic data formats, particularly *property graphs*, is that it applies reference schemas and consistency rules before assigning predicates to statements [5]. Its volume, variety, velocity and veracity place it at the forefront of 'big data' approaches [41,90]. Entries in RDF triple stores are predefined entities, rather than simple text strings, and structured into uni-directional statements [4]. In Wikidata, this is further enhanced by the use of qualifiers to provide additional features of the statements. This structure makes building semantic databases using RDF more difficult and time-consuming than alternative systems, especially *property graphs* [5], but it allows a fully regular representation of statements in knowledge graphs where subjects, predicates and objects are standardized and semantically described. Avoidance of typos and synonyms of string-based systems then allows far faster and more precise information retrieval and usage [4].

In the context of the COVID-19 pandemic, an ontological database representing many aspects of the SARS-CoV-2 outbreak has been represented in Wikidata, building on pilot work that was started at the onset of the Zika pandemic [28] and led to the formation of WikiProject Zika Corpus.[5] This Zika project – itself inspired by dedicated Wikiprojects for Medicine[6] and for Source Metadata[7] – laid many of the foundations for the current COVID-19 work in managing fast-changing information: it developed, documented and refined sets of SPARQL queries about an ongoing epidemic, the underlying pathogen, the disease and diagnostic or therapeutic options, and it piloted workflows for integrating distributed knowledge from multiple databases to build a consistent semantic representation of a topic for which relevant concepts were often not yet readily available through formal ontologies.

Wikidata is apt to cover gaps in ontologies, as any user is entitled to create new classes and propose new properties. In contrast to DBpedia, which is based on scheduled scraping of Wikipedia, the openness of the Wikidata data model allows flexible, immediate representation by any stakeholder interested on a subject. For example, Wikidata has a concept for *vaccine candidate* [Q28051899], but as it is not on a particular Wikipedia page, the concept has never been represented on DBpedia.

The core of the COVID-19 knowledge graph in Wikidata is formed by three main items (red in Fig. 2): *COVID-19* [Q84263196], *SARS-CoV-2* [Q82069695], and *COVID-19 pandemic* [Q81068910]. Those three core COVID-19-

[4] https://web.archive.org/web/20210620223556/https://dbpedia.org/page/Ahmed_Al-Qadri.
[5] https://www.wikidata.org/wiki/Wikidata:WikiProject_Zika_Corpus.
[6] https://www.wikidata.org/wiki/Wikidata:WikiProject_Medicine.
[7] https://www.wikidata.org/wiki/Wikidata:WikiProject_Source_MetaData.

CORRECTED PROOF

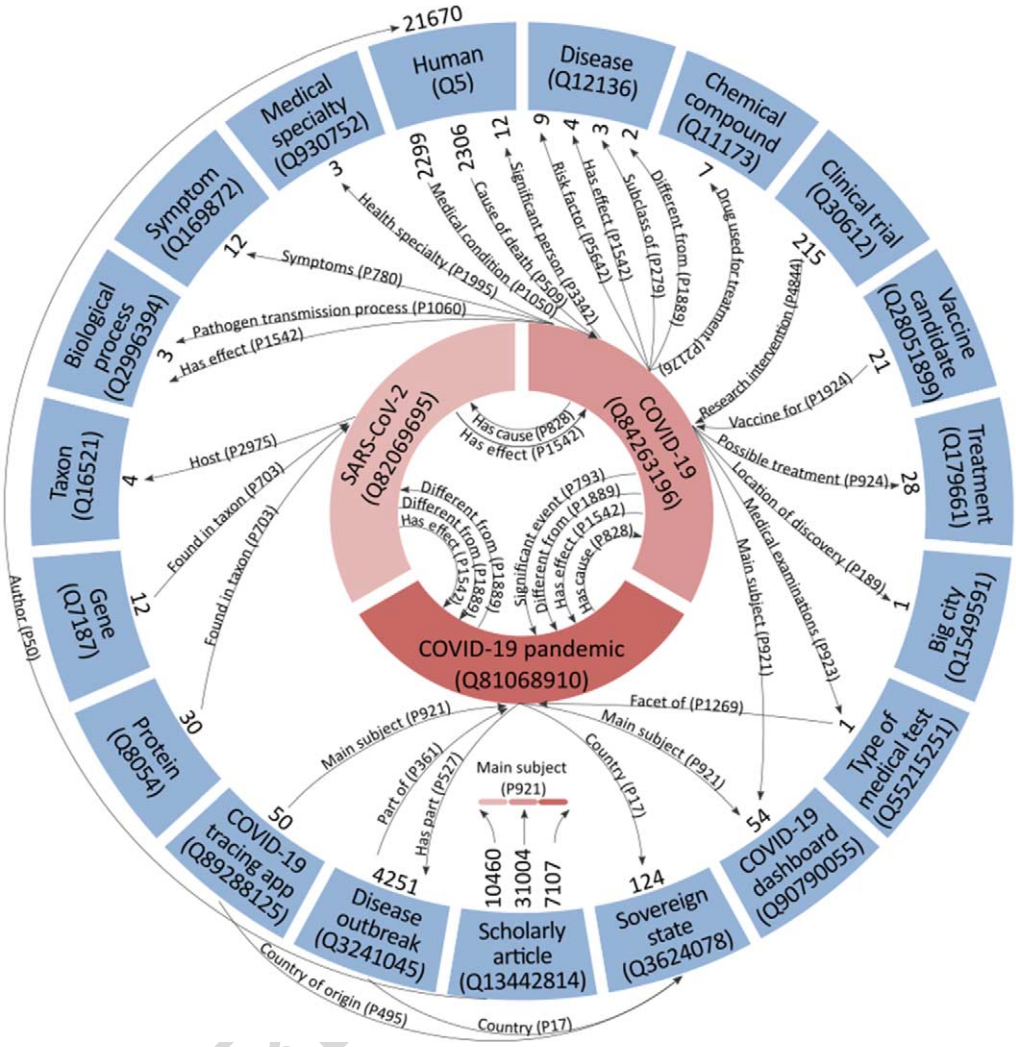


Fig. 2. Simplified skeleton of the data model of COVID-19 information on Wikidata. The three main COVID-related items (the 'C3 items')[8] are represented in red, selected classes of items related to these are shown in blue, with the relations between them represented as arrows. The number of statements relating to each item from the relevant class is indicated next to the item (In the case of scholarly articles, relations to each of the three COVID-related items is indicated by colour). Relation types regularly used to define items within Wikidata classes are omitted (e.g. *chromosome* [P1057] for human genes), as of 20 August 2020,[9] available at: https://w.wiki/auD, license: CC BY 4.0.

related Wikidata items have relatively simple links to one another. Mainly that SARS-CoV-2 causes COVID-19, which itself has had the downstream effect of the COVID-19 pandemic.

These three core items then link to a vast array of items related to all aspects of the disease, its causative virus, and the resulting pandemic (>17,000 Wikidata items as of 20 August 2020; blue in Fig. 2).

When comparing the number of COVID-related Wikidata items with the number of COVID-related entries on the English DBpedia as of May 26, 2021, we find that only 8727 DBpedia entities have been defined for COVID-19 information, presumably only the entities having a corresponding article in English Wikipedia.[10] The limitation of DBpedia in this context is explained by the narrower notability criteria used in Wikipedia for deciding whether a topic can have a Wikipedia article or not. This criterion disallows multiple aspects of COVID-19 from being included in Wikipedia and consequently in DBpedia [32].

[10]For an update of the number of COVID-related items in English DBpedia, please see https://tinyurl.com/ftn5aef3.

The collaborative work in Wikidata to populate and curate this data has been largely accomplished by WikiProject COVID-19,[11] launched in March 2020 [105]. This WikiProject itself has a Wikidata item [Q87748614], and items are linked to it using the property *on focus list of Wikimedia project* [P5008]. The Wikipedia wing of the project similarly works to enrich and validate COVID-19 information through manual fact-checking from referred scholarly publications and trustworthy mass media to prevent vandalism [9] and verify the coverage of the disease through Wikidata-driven analysis of the online encyclopedia [56]. This replication mainly serves to ensure the manual extraction of accurate COVID-19 information from Wikipedia to further enrich Wikidata by contrast to DBpedia that gives less interest to the verification and vandalism identification of Wikipedia information before automatically retrieving it to enrich DBpedia with updated statements [32].

These COVID-19-related items are linked to their respective classes or types using *instance of* [P31] or *subclass of* [P279] relations, and they are linked between each other using non-taxonomic relations defining knowledge about various and multi-disciplinary aspects of COVID-19 (Fig. 2). Biomedical relations between Wikidata items can be assigned *nature of statement* [P5102] or *sourcing circumstances* [P1480] qualifiers to state the status (e.g. official, hypothesis and de facto) and the occurrence probability (e.g. rarely, possibly and often) of the described semantic relation. The network of these items and relations forms a large-scale knowledge graph for COVID-19, where the three core COVID-19-related items noted above extensively link various classes, most notably: *disease outbreaks* [Q3241045] in regions such as continents, sovereign states, and constituent states, *COVID-19 tracing apps* [Q89288125], *COVID-19 vaccines* [Q87719492] and *vaccine candidates* [Q28051899], *scholarly articles* [Q13442814] and *COVID-19 dashboards* [Q90790055]. This graph with short paths to the core COVID items is augmented by biomedical, geographical and other more distantly related entities that are already available in Wikidata, representing an important overview of clinical and other knowledge [98,105]. Such distantly related entities are also available in other open knowledge graphs, particularly DBpedia and YAGO, and contribute much to the value of a semantic resource [30,84]. In Wikidata, several initiatives such as WikiCite for scholarly information [67,70,92,107] and Gene Wiki for genomic data [11] have enabled COVID-19 knowledge graphs to include classes like *genes* [Q7187], *proteins* [Q8054] or *biological processes* [Q2996394], along with the definition of semantic relations between items closely and distantly related to COVID-19. This, consequently, allows the expansion of the coverage of COVID-19 information in Wikidata and a better characterization of COVID-19-related items.

In addition to relational statements that link items within the knowledgebase, non-relational statements link to external identifiers or numerical values [27]. Wikidata items are assigned their identifiers in external databases, including semantic resources, using human efforts and tools such as Mix'n'match [65]. These links make Wikidata a key node of the open data ecosystem, not only contributing its own items and internal links, but also bridging between other open databases (Fig. 3). Wikidata therefore supports alignment between disparate knowledge bases and, consequently, semantic data integration [11] and federation [65] in the context of the linked open data cloud [21]. Such statements also permit the enrichment of Wikidata items with data from external databases when these resources are updated, particularly in relation with the regular changes of the multiple characteristics of COVID-19. By contrast, DBpedia mainly uses Wikipedia for its enrichment and this does not support the coverage of multiple aspects of the analyzed disease [1,45,60]. Examples of Wikidata properties used to define external identifiers can be found in Table 1.

Numerical statements are assigned to disease outbreak items for the COVID-19 pandemic to outline the evolution of the epidemiological status of different entities, from countries to provinces, cities and cruise ships. The properties used to define these statistical statements are shown in Table 1 and include data about the morbidity, the mortality, the testing and the clinical management of COVID-19 at the level of continents, countries and constituent states and also many smaller entities. Some Wikidata properties used to store this epidemiological information have been created in response to COVID-19 (e.g. *Number of recoveries* [P8010], *number of clinical tests* [P8011], and *number of hospitalized cases* [P8049]) proving the flexibility of the knowledge base. To keep records of the progress of the COVID-19 pandemic over time, each statistical statement is assigned a *point in time* [P585] relation as a qualifier. These epidemiological statements are retrieved from CC0 databases such as the COVID-19 DataHub database[12]

[11] https://www.wikidata.org/wiki/Wikidata:WikiProject_COVID-19.
[12] https://datahub.io/core/covid-19.

CORRECTED PROOF

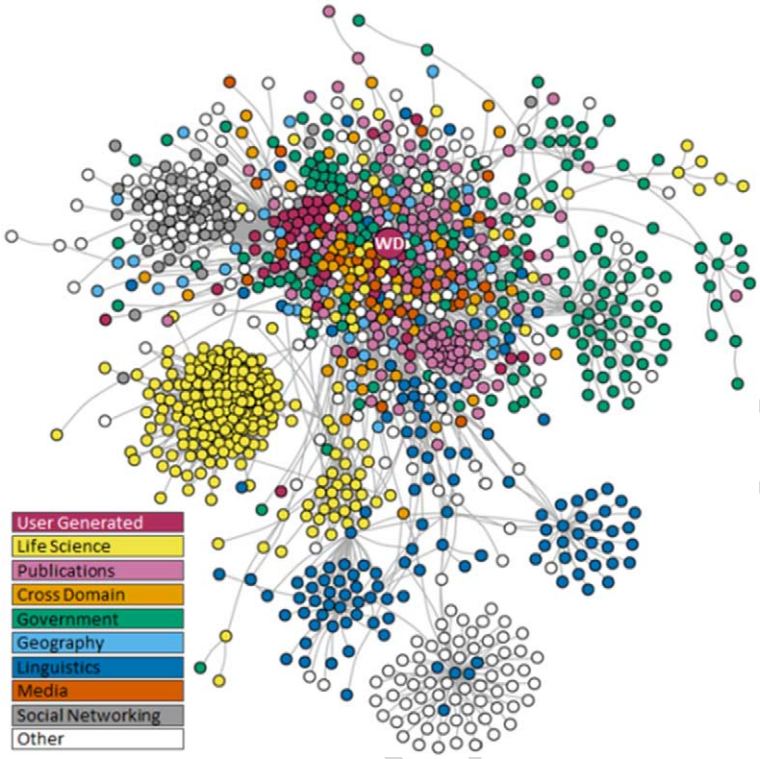


Fig. 3. Wikidata in the Linked Open Data Cloud. Databases indicated as circles (with Wikidata indicated as 'WD'), with grey lines linking databases in the network if their data is aligned, source dataset last updated May 2020 (available at: https://w.wiki/bYM, license: CC BY 4.0).

CORRECTED PROOF

and are linked to them as references. These statements can be used to automatically infer other measures that are not supported by Wikidata but give a full overview of the epidemiology of COVID-19: let c be the total number of confirmed cases at a given day Z when the epidemiological evaluation takes place, d the number of confirmed deaths until that day, r the number of confirmed recoveries by that day, h the number of confirmed hospitalized cases on that day, t the number of clinical tests until that day. On the basis of these values (which could all be represented in Wikidata if matters related to the multi-level coverage of COVID-19 knowledge and conflicts of information from multiple sources are solved), the following measures can be inferred:

- Confirmed active cases $v = c - (d + r)$;
- Confirmed recovery rate $a = r/c$;
- Confirmed patient-days $p = \sum h$ if all infection days are represented;
- New confirmed cases $nc_Z = c_Z - c_{Z-1}$;
- New confirmed deaths $nd_Z = d_Z - d_{Z-1}$;
- New clinical tests $nt_Z = t_Z - t_{Z-1}$;
- New confirmed recoveries $nr_Z = r_Z - r_{Z-1}$.

This set of COVID-19 information is integrated into Wikidata using human efforts, the QuickStatements tool,[13] the Wikidata API,[14] and bots mainly written in Python (e.g. CovidDatahubBot[15]), which explains its quantity and coverage [98]. Data validation is accomplished at multiple layers:

[13] QuickStatements (QS) is a web service that can modify Wikidata, based on a simple text commands: https://quickstatements.toolforge.org/.

[14] An application programming interface (API) is a machine-friendly interface of a web service that can be used to feed another computer program with needed information. The Wikidata API is available at https://www.wikidata.org/w/api.php.

[15] https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/CovidDatahubBot.

Table 1
Examples of Wikidata properties used to define non-relational statements

Wikidata ID	Name	Description
Properties for the alignment with scholarly databases		
P496	ORCID iD	identifier for a researcher (Open Researcher and Contributor ID)
P1153	Scopus Author ID	identifier for an author in the Scopus bibliographic database
P214	VIAF ID	identifier for the Virtual International Authority File database
P7859	WorldCat Identities ID	entity on WorldCat for authority control of authors' data
P1053	ResearcherID	identifier for a researcher in a system for scientific authors, primarily used in Web of Science
Properties for the alignment with clinical language resources and encyclopedias		
P494	ICD-10	identifier in the ICD catalogue codes for diseases – Version 10
P672	MeSH tree code	Medical Subject Headings (MeSH) codes are an index and thesaurus for the life sciences (≠ MeSH ID, P486)
P1417	Encyclopædia Britannica Online ID	identifier for an article in the online version of Encyclopædia Britannica
P486	MeSH descriptor ID	identifier for Descriptor or Supplementary concept in the Medical Subject Headings controlled vocabulary
P3098	ClinicalTrials.gov Identifier	identifier in the ClinicalTrials.gov database
P6680	MeSH term ID	identifier of a "MeSH term" (Medical Subject Headings)
P6694	MeSH concept ID	identifier of a Medical Subject Headings concept
Properties for the non-relational characterization of Wikidata items		
P569	date of birth	date on which the subject was born
P856	official website	URL of the official homepage of an item (current or former)
P1603	number of cases	cumulative number of confirmed, probable and suspected occurrences
P1120	number of deaths	total (cumulative) number of people who died since start as a direct result of an event or cause
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 (equal to the quotient of the number of cases by the number of deaths as stated in a given day)
P8010	Number of recoveries	number of cases that recovered from disease
P8011	number of clinical tests	cumulative number of clinical tests
P8049	number of hospitalized cases	number of cases that are hospitalized
P3488	minimal incubation period in humans	minimal time between an infection and the onset of disease symptoms in infected humans
P3487	maximal incubation period in humans	maximal time between an infection and the onset of disease symptoms in infected humans
P3492	basic reproduction number	number of infections caused by one infection within an uninfected population

CORRECTED PROOF

– Wikidata properties can have constraint declarations associated with them which represent conditions on the use of those properties. As an example, property drug or therapy used for treatment [P2176] has a type constraint that states that the items described by it should be instances of health problem [Q2057971] and a value-type constraint that states that the referenced items should be instances of medication [Q12140].
– In 2019, Wikidata added a new namespace to define Entity schemas using the Shape Expressions (ShEx) language [78,97]. Entity schemas can be used to define expectations about the topology associated with some entities. Entity schemas are human readable and machine processable, facilitating their creation by domain experts and their use for validation. During the pandemic, entity schemas related to COVID-19 entities were created like virus taxon [E192], strain [E174], disease [E69], virus strain [E170], virus gene [E165], coronavirus pandemic local outbreaks [E188] and so on [105]. A remarkable aspect of entity schemas in Wikidata is their collaborative nature, which allows the entity schema ecosystem to evolve by users creating new schemas with different constraints or reusing existing schemas by importing existing ones.
– Another approach to validate the data has been the use of SPARQL queries. SPARQL is available as part of the Wikidata Query Service and it can be used not only to query the knowledge graph, but also to detect

inconsistencies and check logical constraints and more complex patterns based on heuristics. It is also possible to check the edit history and use the ORES service to eliminate database vandalism [97].

Although Web Ontology Language (OWL) can define knowledge graphs with a richer semantic characterization of data models by providing a layer of Description Logics such as in DBpedia [1], the infrastructure developed for the validation of RDF data in Wikidata helps assure a high level of consistency of the Wikidata knowledge graph.

In the context of COVID-19, numerical statements related to epidemiology are constantly changing, and Wikidata's structure benefits it in terms of recency.[16] Wikidata's speed can be tailored by independent users to be as fast as needed. In the case of COVID-19, while death counts on Wikidata might lag behind Wikipedia, bots and humans fill the data model directly, and do not wait for update cycles. As of June 2, 2021, DBpedia counts less than 160.000 deaths in Brazil[17] while both English Wikipedia and Wikidata present counts of over 460.000.[18] This is mainly due to the long delays between the regular updates of the DBpedia statements from Wikipedia. In fact, as of June 2, 2021, DBpedia information for COVID-19 pandemic in Brazil is inferred from the edition of October 30, 2020,[19] leaving it currently outdated. Even if DBpedia has been timely updated from infoboxes, the information infoboxes include can remain outdated in several languages due to the lack of activity on a particular page. An example of such a limitation is the epidemiological data in the infobox of the article about COVID-19 pandemic in France in Simple English Wikipedia as of May 28, 2021 that shows outdated values of confirmed cases and deaths going back to May 13, 2021.[20] In that direction, the Wikidata Bridge project[21] aims to develop Wikidata-driven infoboxes that can be editable in Wikipedia. That may benefit non-English Wikipedia projects in particular to have updated information in infoboxes.

2.2. *Broad multilingual representation*

Wikidata's language-independent data model makes it well-adapted for multilingual representation. In English, French, German and Dutch, its biomedical language coverage is comparable to other semantic resources such as SNOMED-CT,[22] BabelMeSH,[23] and ICD-10[24] (largely due to links with other biomedical ontologies and knowledge bases) [98]. However, its coverage over a larger lange set is markedly broader than those resources [39,61]. Looking at the coverage of the 17,000 items that link to and from the three main COVID items (Fig. 2) and the 55,000 statements that relate them, distinct patterns emerge (Fig. 4E). More than 40% of the predicates (Curves B and D) and more than 90% of the objects (Curve C) of the statements related to COVID are represented in fifty languages or more. English is the unsurprising front-runner in items with COVID as the object, since many of those items are journal articles with untranslated titles (Fig. 4A). The names of the properties that link them (Fig. 4B, D) have much more even coverage, as do items with COVID as the subject (Fig. 4C). This broad but uneven coverage is particularly relevant for multilingual resources for natural language processing purposes in clinical contexts [20].

The better coverage in English is explained in part by the higher support of this language in both biomedical language resources [34] and Wikipedia [91]. Cooperation with publishers such as Cochrane has a significant effect

UNCORRECTED PROOF

[16] This can be illustrated, for instance, by a query for the number of COVID-related items that have been last modified in May and June 2021 (as of June 21, 2021): https://w.wiki/3XBL.

[17] https://web.archive.org/web/20210602140025/https://dbpedia.org/page/COVID-19_pandemic_in_Brazil.

[18] https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic_in_Brazil&oldid=10260573, https://www.wikidata.org/w/index.php?title=Q86597695&oldid=1433353285.

[19] https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic_in_Brazil&oldid=986289428.

[20] https://simple.wikipedia.org/w/index.php?title=COVID-19_pandemic_in_France&oldid=7562853.

[21] https://www.mediawiki.org/wiki/Wikidata_Bridge.

[22] http://www.snomed.org/snomed-ct/sct-worldwide (Accessed February 3, 2021): SNOMED-CT supports English, French, Danish, Dutch, Spanish, Swedish, and Lithuanian.

[23] https://lhncbc.nlm.nih.gov/project/babelmesh-and-pico-linguist (Accessed on February 3, 2021): BabelMeSH supports Arabic, Chinese, Dutch, English, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish, and Swedish.

[24] ICD-10: International Classification of Diseases, 10th Revision [49]: ICD-10 supports Arabic, Chinese, English, French, Russian, Spanish, Albanian, Armenian, Azeri, Basque, Bulgarian, Catalan, Croatian, Czech, Danish, Dutch, Estonian, Persian, Finnish, German, Greek, Hungarian, Icelandic, Italian, Japanese, Korean, Latvian, Lithuanian, Macedonian, Mongolian, Norwegian, Polish, Portuguese, Serbian, Slovak, Slovenian, Swedish, Thai, Turkish, Turkmen, Ukrainian, and Uzbek.

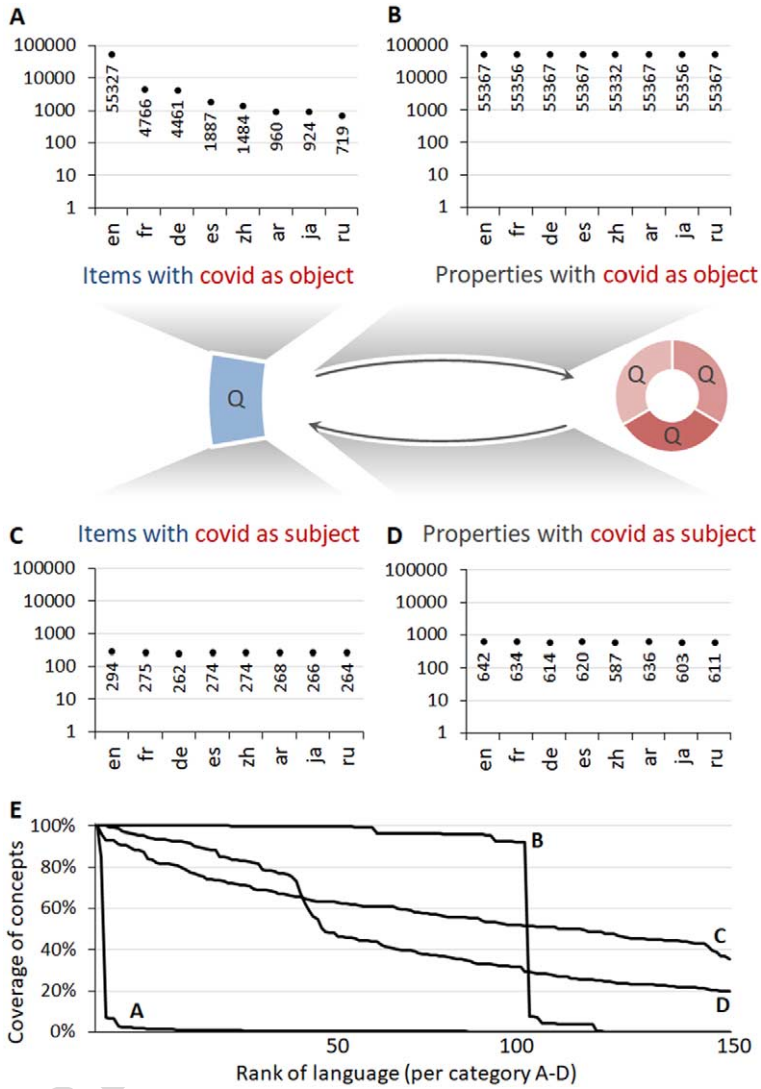


Fig. 4. Language representation of COVID-19-related statements. A-D) Language coverage for items and properties used in statements when either the object or subject is one of the three COVID-related items (as per Fig. 2; note: log y-axis). The eight most common languages in Wikidata are shown: en=English, fr=French, de=German, es=Spanish, zh=Chinese, ar=Arabic, ja=Japanese, ru=Russian.) E) Percentage of the items covered in order from highest to lowest coverage. faceted by categories A-D. Data shown for top 150 languages in each category (note: languages not necessarily in same order for each), as of August 15, 2020 (available at: https://w.wiki/auE, license: CC BY 4.0; live data: https://w.wiki/Yj$, https://w.wiki/Yk3, https://w.wiki/Yk5, and https://w.wiki/Yk6).

on English Wikipedia coverage, too [48]. The significant coverage of languages like French, Spanish, German, Chinese and Swedish in Medical Wikidata fits with their support by major biomedical multilingual databases: ICPC-2 [86] supports 24 languages,[25] SNOMED-CT supports 7 languages, LOINC[26] supports 13 languages, BabelMeSH [33] supports 13 languages, and ICD-10 supports 42 languages. The situation is even worse for other multilingual

[25] ICPC-2 supports Afrikaans, Basque, Chinese, Croatian, Danish, Dutch, English, Finnish, French, German, Greek, Hebrew, Hungarian, Italian, Japanese, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovenian, Spanish, and Swedish.

[26] https://loinc.org/international/ (Accessed on August 13, 2020): LOINC supports Chinese, Dutch, Estonian, English, French, German, Greek, Italian, Korean, Portuguese, Russian, Spanish, and Turkish.

Table 2

Languages ranked by medical content from the literature: number of medical Wikipedia articles, number of Wikidata labels, number of native speakers, and number of Wikidata users. Style code: italic for languages appearing in all four lists; bold for those appearing in only one

Rank	Medical Wikipedia, 2013 [38]		Wikidata labels, 2017 [52]		Population, 2019 [26]		Wikidata users, 2018 [51]
	Language	Number of medical articles	Language	Rate of labels	Language	Native speakers (millions)	Language
1	*English*	29072	*English*	11.04%	**Chinese**	1323	*English*
2	German	7761	Dutch	6.47%	*Spanish*	463	French
3	French	6372	French	6.02%	*English*	369	German
4	*Spanish*	6367	German	5.08%	**Hindi**	342	*Spanish*
5	**Polish**	5999	*Spanish*	4.07%	**Arabic**	335	Italian
6	Italian	5677	Italian	3.9%	Bengali	228	*Russian*
7	Portuguese	5269	**Swedish**	3.89%	Portuguese	227	Dutch
8	*Russian*	4832	*Russian*	3.54%	*Russian*	154	Japanese
9	Dutch	4391	**Cebuano**	2.21%	Japanese	126	**Danish**
10	Japanese	4303	Bengali	1.94%	**Western Punjabi**	82.5	Portuguese

open knowledge graphs, particularly DBpedia which failed to develop COVID-19 information in major languages like French and German as of June 20, 2020.[27]

The support of other natural languages can also be explained by the use of bots that extract multilingual terms representing clinical concepts based on natural language processing techniques and machine learning[28] [93] and by the involvement of research institutions and scientists speaking these languages, particularly German and Dutch, in adding biomedical information to Wikidata [79,103]. The near-100% coverage for properties with COVID-19 as the subject in the most spoken languages (Fig. 4B) resulted from early systematic volunteer translation drives for common properties by *WikiProject Labels and Descriptions*[29] and others [88]. Language coverage of medical Wikidata labels (particularly for diseases' class) seems influenced by several factors. Most obvious for a collaborative project is the number of speakers of each language among the contributor community [51]. However, there also appears to be an impact from the overall number of Wikidata labels for each language [52] and to the number of medical Wikipedia articles in each language [38] (Table 2).

These correlations can be interrogated by querying Wikidata to find out the current status of the editing of this knowledge graph and of Wikipedia in 307 languages (Table S3; top-ranking items for each variable summarised in Tables 3 and 4). Query results largely match previously published trends for Wikipedia and Wikidata (Table 2), though we note that Arabic (ar) and Chinese (zh), appear in the top 10 languages in the Wikidata COVID-19 subset, while being absent from the top 10 s for other sets described in Table 4. Coverage differed across languages and variables, and most of the distributions showed marked positive skew. Nonparametric analysis of correlations (Spearman's rho) found large magnitude associations (rho 0.65 to 0.97, median = 0.84, Supplementary Table S4), statistically significant even following stringent Bonferroni correction. To account for skew and data spanning multiple orders of magnitude, log 10-transformed data was used for subsequent analyses. Pearson's correlation coefficients between all variables was high (Fig. 5). A principal component analysis for the 90 languages with complete data on all 7 indicators found that a single component explained 81% of the variance, with loadings ranging from 0.80 to 0.95. The smallest PCA loading and Spearman's correlation were for the number of viewers, which though still a strong association, was less correlated than the other variables by a substantial margin.

Similarly, the current representation of COVID-19 Wikidata items in natural languages seems to be linked with COVID-19-related Wikipedia pages, edits and pageviews for a given language, as shown in Table 4. This is con-

[27] Please refer to https://web.archive.org/web/20210620132047/http://fr.dbpedia.org/page/COVID-19 and https://web.archive.org/web/20210620132401/http://de.dbpedia.org/page/COVID-19 as a proof of this assumption.

[28] An example of such a Wikidata bot can be Edoderoobot 2, which is specifically working on labelling, thereby translating structured data into prose in the respective language. Further information about this bot can be found at https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/Edoderoobot_2.

[29] https://www.wikidata.org/wiki/Wikidata:WikiProject_Labels_and_descriptions.

CORRECTED PROOF

Table 3

Languages ranked by medical content from Wikidata queries (as of August 11, 2020). The medical Wikipedia query yields Wikipedia articles associated with Wikidata items that have a disease ontology ID [P699] or are in the tree of any of the following classes: medicine [Q11190], disease [Q12136], medical procedure [Q796194] or medication [Q12140]. The medical Wikidata labels query yields labels of Wikidata items that have a disease ontology ID [P699] or a MeSH descriptor ID [P486] or are in the tree of any of the same four classes. The Wikipedia and Wikidata users column provides a snapshot from the Wikidata dashboard that lists Wikidata users who also edit Wikipedia by number of such users per Wikipedia language. Style code: italic for languages appearing in all three lists; bold for those appearing in only one

Rank	Medical Wikipedia articles https://w.wiki/Z6a		Medical Wikidata labels https://w.wiki/Z6h		Wikipedia and Wikidata users https://w.wiki/Z6W	
	Language	Number of medical articles	Language	Number of labels	Language	Number of users
1	*English*	16670	*English*	65986	*English*	9600
2	*German*	8911	*French*	37053	*French*	2580
3	*Arabic*	8596	*German*	22432	*German*	2490
4	*French*	7258	*Spanish*	21505	*Spanish*	2330
5	*Spanish*	6979	*Arabic*	18581	*Russian*	1790
6	*Italian*	6498	*Italian*	18074	*Italian*	1430
7	**Polish**	6071	*Japanese*	17992	Chinese	1120
8	Portuguese	5652	**Dutch**	17985	*Japanese*	1090
9	*Russian*	5564	Chinese	17462	Portuguese	979
10	*Japanese*	4651	*Russian*	17165	*Arabic*	688

CORRECTED PROOF

firmed by the high correlation (Pearson $r = 0.93$) of the language distribution of COVID-related Wikidata labels with the number of COVID Wikipedia pages in language editions and the moderate correlation (Pearson $r > 0.65$) between the number of Wikidata COVID-related labels in a given language and the quantity and edit statistics of medical content in Wikidata and Wikipedia (Fig. 5). Such relationships are strengthened by the high correlation (Pearson $r > 0.9$) between the number of medical Wikidata labels in a given language and the number of medical Wikipedia articles in language editions as well as the number of native speakers jointly editing Wikipedia and Wikidata.

To investigate the possible causes of these highly correlated datasets, we compared them to two external metrics for each language: the number of native speakers of each language [26] and the maximum Human Development Index for countries where that language is an official language [99]. This data was available for fewer languages ($N = 57$ each, 19 pairs) and the sparse overlap precluded including both simultaneously in analyses. The number of native speakers showed a similar positive skew to earlier data, so was also log 10-transformed. Even though these analyses are necessarily exploratory, maximum development correlated more strongly than did the number of speakers (Fig. 5B; Table S4). Cohen's q values (an effect size for differences between correlation coefficients) of a size considered unusually large for the social sciences (>0.5) were observed when comparing correlation of development index versus number of speakers with the number of medical Wikidata labels and with the number of users. Further medium q values (differences > 0.3) were observed for correlation to the number of medical Wikipedia articles and to the number of COVID Wikipedia pages. Correlation differences were negligible with regard to development versus number of speakers as associated with the number of edits or pageviews [15].

The observation here that current language coverage in Wikidata and Wikipedia correlates more closely to countries' development index than to the number of speakers of each natural language aligns with previous work demonstrating low correlation of Wikidata to the number of speakers [52].

We interpret this as a potential 'need gap', where languages that have a large number of relatively low-income speakers remain relatively underserved. To address this, it may be necessary to encourage and/or support contribution by speakers of under-resourced and unrepresented languages to medical Wikidata projects, analogous to those Wikipedia projects.[30]

[30]Current efforts to enhance the coverage and language support of medical knowledge in Wikipedia are mainly driven by Wikimedia Medicine. For further information, please refer to https://meta.wikimedia.org/wiki/Wiki_Project_Med. An example of the initiatives under this umbrella is

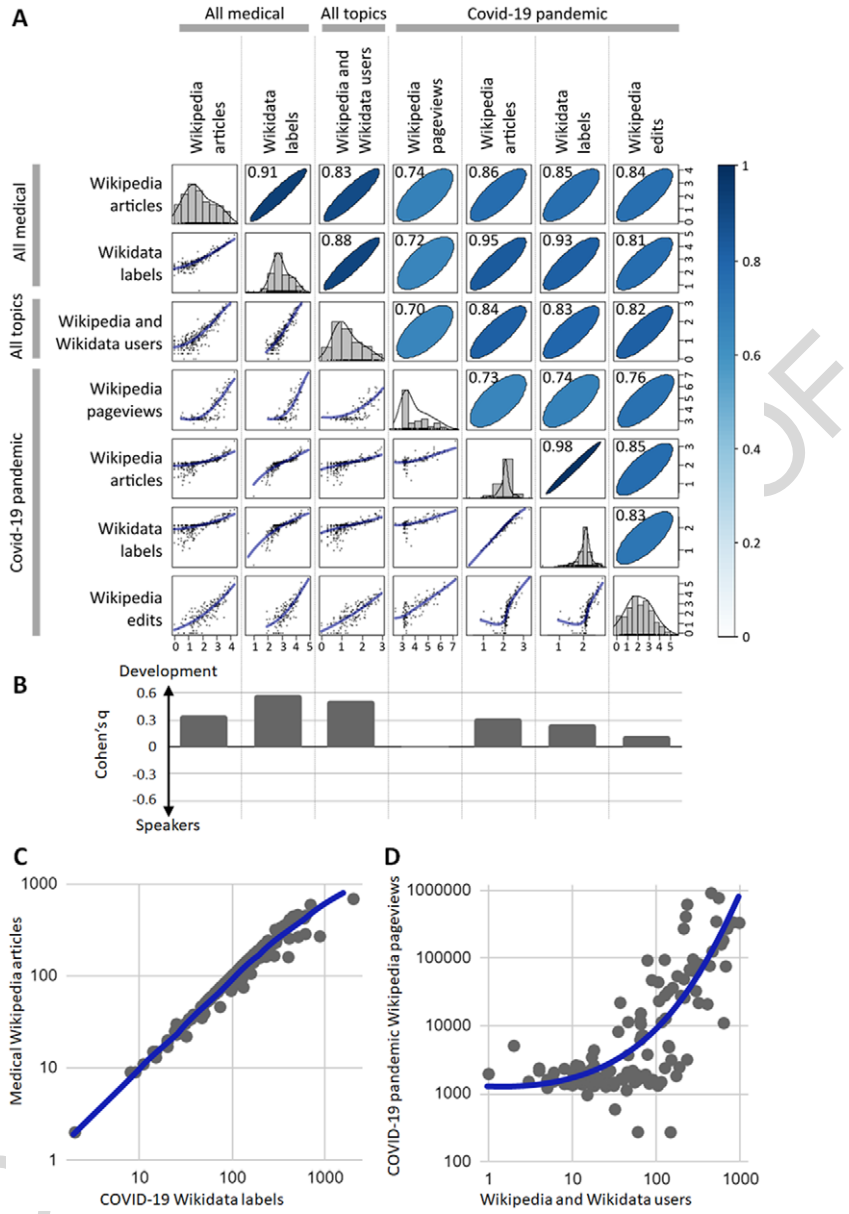


Fig. 5. A) all-versus-all pairwise correlations of log 10-transformed values of seven metrics for 307 languages (data from sources detailed in Tables 3 and 4). Histograms on diagonal indicate skew, scatter plots below diagonal indicate data and trendlines, ellipsoids above diagonal indicate Spearman's R correlation coefficient. b) Cohen's q coefficient comparing correlation of the seven metrics to maximum human development index versus to the number of native speakers. C) highest correlated variable pair. d) lowest correlated variable pair. [Available at: https://w.wiki/zV6, license: cC-BY 4.0].

In addition to the intrinsic value of increased language coverage, it would also help in ensuring culturally relevant contextualizations in Wikidata's medical and other domains.

the Special Wikipedia Awareness Scheme for The Healthcare Affiliates project, focused on languages of India. An explanation of this project can be found at https://en.wikipedia.org/wiki/Wikipedia:SWASTHA.

Table 4

Languages ranked by COVID-19-related content from Wikidata queries and other live data (as of August 13, 2020). The COVID-19 pandemic Wikipedia pageviews column represents daily average user traffic (averaged over 2020) to the article about the COVID-19 pandemic in the respective language. The COVID Wikidata labels query sorts languages by the number of labels of Wikidata items with a direct link to and/or from any of the core COVID-19 items – Q84263196 (COVID-19), Q81068910 (COVID-19 pandemic) and Q82069695 (SARS-CoV-2) – excluding items about humans (3131) or scholarly publications (40164). The COVID Wikipedia articles query filters those Wikidata items for associated Wikipedia articles and sorts languages by the number of such articles. The values in the COVID Wikipedia edits column represent the revision counts per Wikipedia language as taken from the dashboard listing Wikimedia projects by total number of revisions to COVID-19-related articles. Style code: Italic for languages appearing in all four lists; bold for those appearing in only one

Rank	COVID-19 pandemic Wikipedia pageviews https://w.wiki/ZTG		COVID Wikipedia articles https://w.wiki/ZSt		COVID Wikidata labels https://w.wiki/ZSq		COVID Wikipedia edits https://w.wiki/y9u	
	Language	Avg. daily pageviews	Language	Number of articles	Language	Number of labels	Language	Number of edits
1	*English*	52872	*English*	561	*English*	1429	*English*	250306
2	Russian	41246	*Arabic*	517	**Dutch**	785	*German*	126359
3	*Spanish*	37722	*German*	431	*Arabic*	623	*French*	42029
4	*Chinese*	27598	Portuguese	427	**Catalan**	579	*Chinese*	41545
5	*German*	20707	**Korean**	408	*German*	561	*Spanish*	30869
6	**Italian**	8490	*Chinese*	396	*French*	517	*Arabic*	19963
7	*French*	7959	**Vietnamese**	392	Japanese	503	Russian	18719
8	Portuguese	7648	*French*	379	*Chinese*	483	Japanese	11508
9	Japanese	5227	*Spanish*	370	Portuguese	463	**Ukrainian**	10599
10	*Arabic*	4300	**Indonesian**	363	*Spanish*	433	**Hebrew**	10386

2.3. *Alignment to external databases*

As shown in the "Data model" section, Wikidata items are linked to their equivalents in other semantic databases using statements where the property provides details about a given resource and the object is the external identifier of the item in the aligned database. Similarly to Wikidata items, these database alignment properties are defined by labels, descriptions and aliases in various languages and by statements describing logical conditions for their usage including formatting constraints and allowed values of subject classes [97].

The alignment of Wikidata entities to other entries on different databases is a collaborative process which, as with everything in Wikidata, is done via combination of manual and automatic curation. As an example of automation, items concerning scholarly entries (i.e. articles and reports) were often aligned to other databases using DOIs (Digital Object Identifiers) as unique keys. As Wikidata is an open database, the precision of the alignments is largely based on trust in the community, and misalignments are promptly corrected once identified. At the scale of curation happening on Wikidata, quality issues in aligned databases regularly surface, e.g. invalid DOIs stated in PubMed and PMC Europe.[31] While most of these databases have some feedback channels, no mechanisms exist for informing them systematically about issues with their data that have been identified at the scale of Wikidata-based curation.

As of September 1, 2020, 5302[32] out of 7877[33] Wikidata properties are used to state external identifiers of the Wikidata items. These properties facilitate interoperability between Wikidata and other databases and consequently the regular enrichment of Wikidata with detailed information from online ontologies and knowledge graphs updated on a daily basis [29,30,102]. The output using such Wikidata properties can be adapted as an open license framework for the automatic evaluation and learning of knowledge graph alignment approaches [85,102] and for the integration of scholarly knowledge [68].

In the circumstances of the COVID-19 outbreak, a SPARQL query[34] has been formulated to analyze the integration of external identifiers in Wikidata. This query succeeded in returning the main aligned external resources to the set of scholarly articles and clinical trials, of diseases, of symptoms, of drugs, of humans, of sovereign states, of

[31] https://github.com/br2s/bug-reports-to-science/issues/8.

[32] For the updated count of the properties defining external identifiers, refer to https://w.wiki/ayn.

[33] For the updated count of all the properties, refer to https://w.wiki/ayo.

[34] https://w.wiki/auR.

CORRECTED PROOF

genes, of proteins, and of other items related to the ongoing COVID-19 pandemic in Wikidata. This confirms the centrality of Wikidata within the linked open data cloud (cf. Fig. 3 and [21]) and consequently the usefulness of Wikidata to address the COVID-19 data challenges and dynamically integrate various types of semantic data in the context of the disease outbreak.

Scholarly articles and clinical trials have been linked to numerous external identifiers, particularly the Digital Object Identifier (DOI), the PubMed ID, the Dimensions Publication ID, the PubMed Central ID (PMCID) and the ClinicalTrials.gov Identifier (Table S5). Most of these identifiers are added thanks to WikiProject WikiCite aiming to add support of bibliographic information on Wikidata [67,70,107]. The current representation of external identifiers for the scientific literature in Wikidata seems to be similar to the general one for the bibliographic data in the knowledge graph. As of September 3, 2020, 36208373 scholarly articles[35] are currently represented in Wikidata. 31425586 of which have PubMed IDs and 25896956, 6016452, and 346114 scientific publications respectively have DOIs, PubMed Central IDs and ArXiv IDs.

However, this Wikidata coverage of the availability of COVID-19-related publications in external research databases does not seem to fully represent full records of COVID-19 literature in aligned resources. By way of comparison, we performed a simple search for "COVID-19" in a set of literature databases, and there were 103796 COVID-19-related records available on PubMed,[36] 110323 COVID-19 full texts accessible on PubMed Central,[37] 296450 COVID-19 publications on Dimensions,[38] 211000 records on Semantic Scholar,[39] 4778 records at ClinicalTrials.gov,[40] 3295 records on arXiv ID,[41] and 183 records on NIOSHTIC-2[42] as of February 17, 2021.

Wikidata's relatively incomplete coverage of the literature is mainly explained by Wikidata's development of scientific metadata being based on latent crowdsourcing of information from multiple sources through bots and human efforts and not on the real-time screening of the external scholarly resources [92,107]. In addition to such sampling biases, there are also differences in annotation workflows, e.g. in terms of the multilinguality of or the hierarchical relationships between topic tags in Wikidata versus comparable systems like Medical Subject Headings.

As for the diseases and symptoms related to COVID-19, Wikidata maps to multiple external identifiers in the main biomedical semantic databases such as MeSH, ICD-10,[43] and UMLS[44] as well as in open lexical databases like OBO Foundry ontologies (e.g. Human Phenotype Ontology) and Freebase (Table S6). This is mainly due to the use of machine learning algorithms to align these major online biomedical resources to Wikipedia articles and consequently to Wikidata items [81]. The representation of open license resources is particularly explained by the use of these databases to form the core of the biomedical knowledge in Wikidata through mass uploads and timely updates [104]. Items about diseases and symptoms are also aligned to several online encyclopedias (e.g. eMedicine, Encyclopedia Britannica, and MedlinePlus) and to non-medical databases such as scholarly repositories (e.g. JSTOR[45]) and bibliographic databases (e.g. Microsoft Academic[46]) using external identifiers' statements. This can be explained by the efforts of WikiProject Source Metadata[47] and the WikiCite initiative to align topic pages in research databases to Wikidata items, so that active members of this project can easily extract topics of research publications from source databases and assign them to the corresponding Wikidata items using *main subject* [P921] relations [70]. The linking from Wikidata items about between diseases and symptoms to online first-class encyclopedias is not restricted to the context of the COVID-19 pandemic [104] and is a rather established practice to

CORRECTED PROOF

[35] https://scholia.toolforge.org/.
[36] https://pubmed.ncbi.nlm.nih.gov/?term=COVID-19.
[37] https://www.ncbi.nlm.nih.gov/pmc/?term=COVID-19.
[38] https://tinyurl.com/y6kwrdth.
[39] https://www.semanticscholar.org/search?q=COVID-19&sort=relevance.
[40] https://clinicaltrials.gov/ct2/results?cond=COVID-19&term=&cntry=&state=&city=&dist=.
[41] https://arxiv.org/search/?query=COVID-19&searchtype=all&source=header.
[42] https://www2a.cdc.gov/nioshtic-2/Buildqyr.asp?S1=COVID-19&Submit=Search.
[43] International Classification of Diseases, Tenth Revision (https://www.who.int/classifications/icd/en/).
[44] Unified Medical Language System (https://www.nlm.nih.gov/research/umls/index.html).
[45] https://www.jstor.org/.
[46] https://academic.microsoft.com/.
[47] https://www.wikidata.org/wiki/WD:WikiProject_Source.

provide Wikidata users with pointers to further specialized information pertaining to a given Wikidata item [58] and to allow comparison of medical data quality between Wikipedia and other encyclopedias [38].

Since Wikidata is multidisciplinary, it has extensive matching to humans and sovereign states (Table S7) as well as films, computer applications and disease outbreaks (Table S8). The alignment to various metadata databases like VIAF,[48] WorldCat,[49] Library of Congress and IMDb[50] is motivated by the mass import of authority control data for the interoperability between library metadata and for the prevention of the duplication of items including book authors, actors and films [3,58]. Wikidata items about sovereign states and humans are aligned to corresponding topic pages and user pages in social networking services (Twitter) and question answering forums (Quora and Reddit). This enables tracking the effect of the information provided by Wikimedia projects, particularly Wikipedia, on online communities [100]. Information about items in social media can also be retrieved to support the topic modelling of the coverage of the pandemic in social networks [14]. Taken together, these database alignments are useful to integrate new non-clinical information to Wikidata, to allow correlations between epidemiological data and non-medical information about countries, individuals, masterpieces and disease outbreaks such as geopolitical, software programming and economic data, and to provide further readings about the concerned items [68].

Concerning drugs, proteins, genes and taxons, Wikidata items are mainly assigned external identifiers in the major knowledge graphs for pharmacology (e.g. MassBank[51]), for biodiversity (e.g. IRMNG[52]), for genomics (e.g. Entrez Gene) and for proteomics (e.g. PDB[53]) and are rarely linked to non-medical databases or to encyclopedias, as shown in Table S8. The lack of alignment between these biomedical Wikidata items and their equivalents in social web services is explained by the higher interest of social media users in the health policies and epidemiology of COVID-19 rather than the therapeutic options and molecular aspects related to the disease [72]. The most important interest in matching these concepts in Wikidata to graph databases (e.g. Massbank, PDB, and KEGG[54]) and semi-structured databases (e.g. Guide to Pharmacology[55]) for bioinformatics rather than online encyclopedias is due to the better availability of genomic and proteomic information in these specialized semantic resources [43,104]. The alignment of taxon items in Wikidata to biodiversity knowledge graphs (e.g. NCBI[56] taxonomy and IRMNG) is to permit the discussion of the pathogenesis of coronavirus and mainly COVID-19 through the analysis of the physiological features of infected taxons [74]. The sum of these biomedical alignments is developed using human edits and computer tools thanks to large initiatives to develop open ontological databases for curating advanced molecular biology data such as WikiGenomes [79] and Gene Wiki [11] and is enhanced in the context of the current pandemic through the contributions of WikiProject COVID-19 [105].

Despite the volume and variety of database alignment in Wikidata, particularly related to COVID-19, the Wikidata statements providing external identifiers do not provide the extent of matching between the subject and its equivalent in the aligned database. By contrast, DBpedia assigns different properties for database matching according to the level of correspondence between the aligned entities (e.g. rdfs:seeAlso, skos:broader, or owl:sameAs) [94]. As a solution to this matter, a new Wikidata property entitled "mapping relation type" (P4390) has been created. This property is assigned as the predicate to the qualifier of a statement providing an external identifier of an item. The object of this qualifier has to be one of the SKOS generic mapping relation types: "close match" (Q39893184), "exact match" (Q39893449), "narrow match" (Q39893967), "broad match" (Q39894595) or "related match" (Q39894604). When the object is an "exact match", the two aligned items are equivalent. However, when the object is a "broad match", this means that the external entity is a hypernym to the corresponding Wikidata items (i.e. skos:broader), etc.

CORRECTED PROOF

[48] Virtual International Authority File (http://viaf.org/).
[49] https://www.worldcat.org/.
[50] Internet Movie Database (https://www.imdb.com/).
[51] https://massbank.eu/MassBank/.
[52] Interim Register of Marine and Nonmarine Genera (https://www.irmng.org/).
[53] Protein Data Bank (https://www.rcsb.org/).
[54] Kyoto Encyclopedia of Genes and Genomes (https://www.genome.jp/kegg/).
[55] IUPHAR/BPS Guide to Pharmacology (https://www.guidetopharmacology.org/).
[56] National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/).

3. Visualizing facets of COVID-19 via SPARQL

One of Wikidata's key strengths is that each item can be understood by both machines and humans. It represents data in the form of items and statements, which are navigable in a web interface and shared as semantic triples [102]. However, where a computer can easily hold the entire knowledge base in its memory at once, the same is obviously not true for a human.

Since we still rely on human interpretation to extract meaning out of complex data, it is necessary to pass that data from machine to human in an intuitive manner [57]. The main way of doing this is by visualising some subset of the data, since the human eye acts as the input interface with the greatest bandwidth. Because Wikidata is available in the RDF format, it can be efficiently queried using SPARQL,[57] a semantic query language dynamically extracting triple information from large-scale knowledge graphs.

Given the breadth of Wikidata's COVID-19-related information (examples in Supplementary Figure S1), extracting specific subsets of that information using SPARQL[58] can illustrate different aspects of the COVID-19 disease, its causative virus, and the resulting pandemic (Supplementary Table S1). Several Wikidata-based COVID-19 dashboards take advantage of the variety of visualizations that can be generated using the Wikidata Query Service (Supplementary Table S2) from both a quantitative perspective (amount of statistical data that can be generated through the integration of COVID-19 information with non-COVID-19 data) and a qualitative one (visualization types and topics). This section will present examples across different aspects of COVID-19, adapted from five main sources to which we have contributed substantially[59,60,61,62,63] Several similar query collections exist, e.g. for COVID-19 in India[64] and Indonesia [17].

3.1. Biological and clinical aspects

A simple demonstration of Wikidata's encoding of SARS-CoV-2's basic biology is in its genetics (Fig. 6) and resulting symptoms (Fig. 7). The viral genome contains 11 genes that encode 30 proteins (and variants), which are currently known to interact with over 170 different human proteins. Although there are two genome browsers based on Wikidata [66,79], neither yet display the SARS-CoV-2 genome. SPARQL visualizations provide a broader way to explore biomedical knowledge about the studied virus and the related infectious disease. As the knowledge graph grows, this is allowing linking together complex knowledge on biochemistry (e.g. genes and proteins), biology (e.g. host taxa), clinical medicine (e.g. interventions) [104]. Such queries can be expanded by considering the qualifiers that modulate biomedical statements. These qualifiers allow the assignment of weights to assumptions according to their importance and certainty. For instance, some treatments are indicated as hypothetical, or symptoms are listed as rare, as defined by their *nature of statement* [P5102] or *sourcing circumstances* [P1480] qualifiers, with references to back these up (live data: https://w.wiki/bmJ).

CORRECTED PROOF

[57] The recursive acronym for "SPARQL Protocol and RDF Query Language", the current version of which is SPARQL 1.1. A full description of this language is available at https://www.w3.org/TR/sparql11-query/.

[58] Technical documentation about SPARQL can be found at https://en.wikibooks.org/wiki/SPARQL.

[59] WikiProject COVID-19 (WPCOVID) queries: extracts from the query collection of Wikidata's WikiProject COVID-19; https://www.wikidata.org/wiki/Wikidata:WikiProject_COVID-19/Queries.

[60] SARS-CoV-2-Queries: extracts from the book "Wikidata Queries around the SARS-CoV-2 virus and pandemic" [2]; https://egonw.github.io/SARS-CoV-2-Queries/.

[61] SPEED queries: extracts from the Wikidata-based epidemiological surveillance dashboard for COVID-19 pandemic in Tunisia (https://speed.ieee.tn). It was partially built upon COVID-19 Wikidata dashboard (https://sites.google.com/view/covid19-dashboard).

[62] Scholia queries: queries underlying COVID-19-related visualizations from the Wikidata-based scholarly profiling tool Scholia [83]; https://scholia.toolforge.org/.

[63] Covid-19 Summary queries: queries visualizing COVID-19 information in Wikidata linked to the epidemiological information of the outbreak and to the characteristics of the infected famous people; https://public.paws.wmcloud.org/User:99of9/Covid-19.ipynb.

[64] https://w.wiki/LsK.

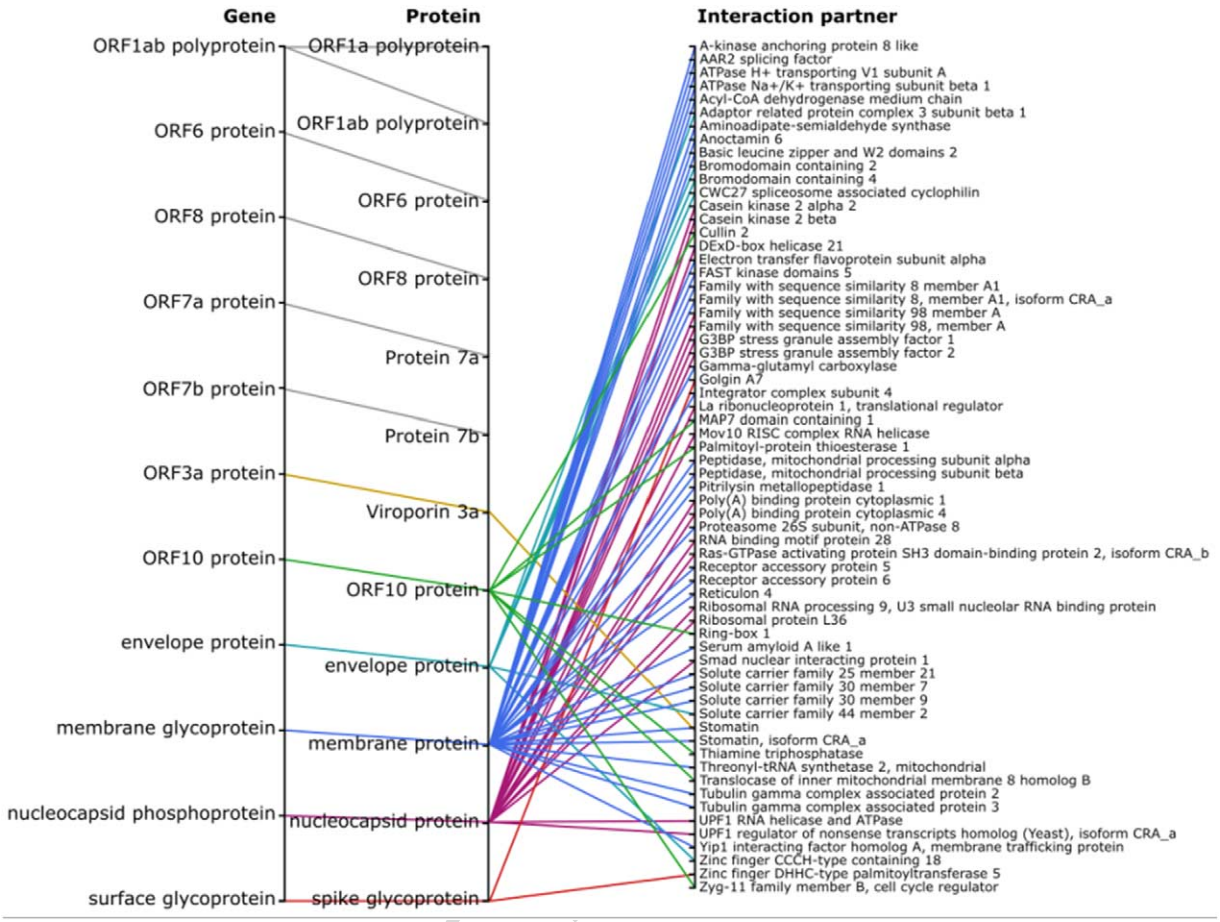


Fig. 6. SARS-CoV-2 interactions with the human proteome as of September 14, 2020 (available at: https://w.wiki/c3D, license: CC BY 4.0). Proteins encoded by SARS-CoV-2 genes (note that some genes encode multiple proteins) and the currently known human protein interaction partners (live data: https://w.wiki/beR).

3.2. *Epidemiology*

Wikidata also contains the necessary information to calculate common epidemiology data for different countries, such as mortality per day per capita, and case number to mortality rate correlation. In some cases this is stored as aggregate data, such as the *case mortality rate* [P3457] statements for regional epidemics stored as numeric data (Fig. 8A), whereas other common visualisations can be calculated from granular data such as the individual *date of birth* [P569] and *date of death* [P570] of notable individuals deceased from COVID-19 (Fig. 8B). Although this reflects the age distribution of COVID mortality, it is also influenced by the demographics of persons sufficiently notable to have Wikidata items.

In some cases, summary data is also time-resolved, allowing inquiry of its change over time (Supplementary Figure S2), capturing features not depicted in several statistical predictions of the epidemiological evolution of COVID-19 outbreaks [12] and clearly seen in other data sources, such that mortality peaks at the beginning of a disease outbreak [111]. Wikidata's granularity (i.e. the representation of COVID-19 information at the scale of individual cases, days and incidents) and collaborative editing have also made it highly up to date on queries such as the most recent death of notable persons due to COVID-19. This result is difficult to achieve with other datasets (Supplementary Figure S3), and mirrors Wikipedia's well-known rapid response to updating information on deaths [55,56].

A



B

disease	diseaseLabel	symptom_count	symptoms
wd:Q21396183	arsenic pentoxide exposure	12	headache // abdominal pain // brain diseases // respiratory failure // cough // dyspnea // nausea // fever // anorexia // diarrhea // delirium // conjunctivitis
wd:Q706845	Lassa fever	10	headache // fatigue // cough // abdominal pain // nausea // brain diseases // fever // myalgia // diarrhea // conjunctivitis
wd:Q21173341	cadmium dust exposure	9	headache // anemia // cough // nausea // dyspnea // chills // anosmia // myalgia // diarrhea
wd:Q21173343	cadmium oxide exposure	9	headache // anemia // cough // nausea // dyspnea // chills // anosmia // myalgia // diarrhea
wd:Q51993	Ebola hemorrhagic fever	8	headache // nausea // dyspnea // fever // myalgia // diarrhea // conjunctivitis // abdominal pain
wd:Q21167939	benzene exposure	8	headache // fatigue // abdominal pain // nausea // dyspnea // respiratory failure // anorexia // diarrhea
wd:Q21402492	colchicine toxicity	8	anemia // abdominal pain // respiratory failure // nausea // fever // anorexia // diarrhea // delirium

Fig. 7. Symptoms of COVID-19 and similar conditions as of September 10, 2020 (available at: https://w.wiki/byX, license: CC BY 4.0). A) Currently listed symptoms of COVID-19, with qualifiers indicating their frequency. (live data: https://w.wiki/N8f). B) Other medical conditions sorted by the number of shared symptoms with COVID-19. (live data: https://w.wiki/bqV; adapted from https://scholia.toolforge.org/disease/Q84263196).

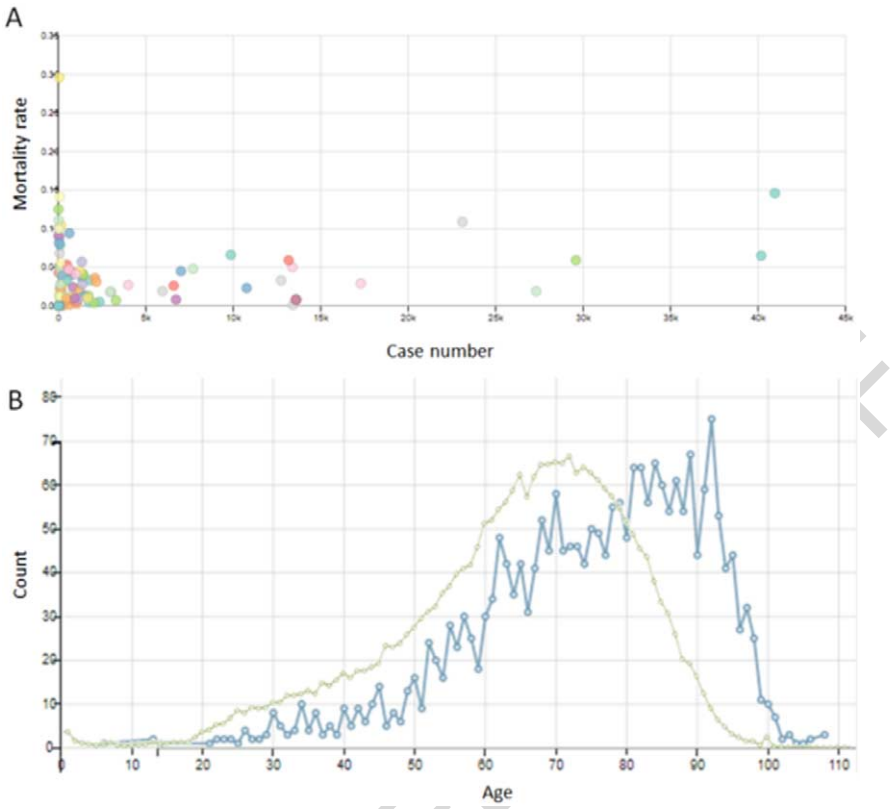


Fig. 8. Summary epidemiological data on the COVID-19 pandemic as of September 10, 2020 (available at: https://w.wiki/byW, license: CC BY 4.0). A) Correlation between the current number of cases and mortality rates in every country, calculated from numeric summary data for each region. Countries coloured randomly (live data: https://w.wiki/bf$). B) Age distribution of notable persons who have died of COVID-19 (blue), compared to the death age distribution for notable persons who were born after 1901 (green), calculated from individual dates of birth and death (live data: https://w.wiki/be7 and https://w.wiki/but).

3.3. Research outputs

A large portion of Wikidata is dedicated to publication metadata and citation links. There are several ways to investigate the relevant topics in publications regarding COVID-19. Firstly, topic keywords can be extracted directly from the titles of articles with COVID-19 as a main topic (Fig. 9A). This is a useful and rapid first approximation of topics covered by those publications, extracted as plain text. These can be expanded upon by querying for the *main subject* [P921] of a set of publications in Wikidata. This property acts analogously to the narrower but more granular Medical Subject Headings (MeSH) descriptors [95]. Such statements allow broader querying of the literature as a network via co-occurrence of topics as the main subject of articles (Fig. 9B).[65] This enables rapid traversal and faceting of the literature on topics in addition to the traditional links made by tracing citations [42], such as extracting common pharmacological and non-pharmacological interventions (live data: https://w.wiki/N8i). The 'WikiCite' project is working on importing the citation network into Wikidata to make a fully open citation network (Fig. S4) [10].

Because Wikidata is agnostic to the exact type of research output, its structure is equally suited to representing information on research publications, preprints (Fig. S5), clinical trials (Fig. S6) or computer applications (Fig. S7). However, preprints are not yet thoroughly covered in Wikidata, a limitation for this context as preprints have become particularly important during the rapid pace of COVID-19 research [10,64]. Further, Wikidata's rich biographical

[65] https://ts404.shinyapps.io/topicnetwork.

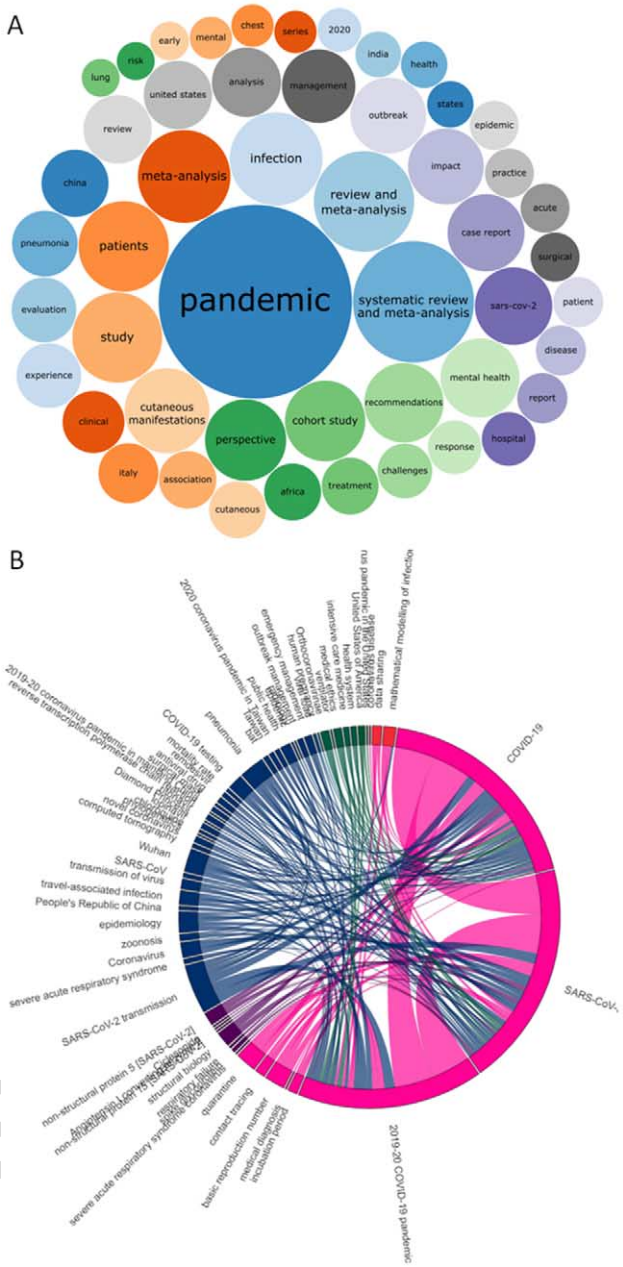


Fig. 9. COVID-19 publication topics as of September 10, 2020 (available at: https://w.wiki/byV, license: CC BY 4.0). A) Common words and word combinations (ngrams) in the titles of publications (live data: https://w.wiki/cFu). B) Co-occurrence of topics in publications with one of the COVID-related items as a topic, with ribbon widths proportional to the number of publications sharing those topics (log scale). Topics coloured by group as determined by Louvain clustering, topics shared in fewer than 5 publications omitted (interactive version: https://csisc.github.io/WikidataCOVID19SPARQL/Fig8B.html; live data: https://w.wiki/bww).

and institutional data makes extracting information on authors, institutions or others straightforward (Fig. S8), and eventually for other contributors too [71].

A

organization	organizationLabel	bankruptcyDate	countryLabel	inception	industries	parents	subsidiaries
wd:Q2208025	STA Travel	20 August 2020	Germany	1 January 1979	tourism industry	DKSH	
wd:Q7606770	Stein Mart	12 August 2020		1 January 1902	retail		
wd:Q5206569	DW Sports Fitness	3 August 2020	United Kingdom	1 January 2009	retail		
wd:Q2749082	Lord & Taylor	2 August 2020	United States of America	1 January 1826	retail		
wd:Q64059182	Le Tote	2 August 2020		1 January 2012	clothing, sharing economy		
wd:Q3305660	Tailored Brands	2 August 2020	United States of America	1 January 1973	retail		Men's Wearhouse
wd:Q15109854	California Pizza Kitchen	30 July 2020	United States of America	1 January 1985	hospitality industry		

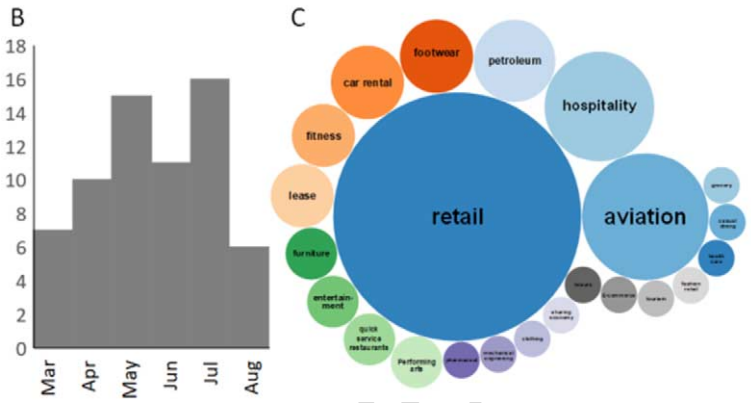


Fig. 10. Bankruptcies of publicly listed businesses due to the COVID-19 pandemic as of September 13, 2020 (available at: https://w.wiki/byY, license: CC BY 4.0). A) Tabular output of SPARQL query B) Bankruptcies per month C) ratios of different industries associated with bankrupt companies. (live data: https://w.wiki/cG6).

3.4. Societal aspects

Further emphasising the multidisciplinary nature of Wikidata, there are also significant social aspects of the pandemic contained in the knowledge base. This includes simple collation of information, such as regional official COVID websites, and unofficial but common hashtags (Fig. S9), or relevant images under Creative Commons licenses (Fig. S10). It also includes more cross-disciplinary information, such as companies that have reported bankruptcy, with the pandemic recorded as the main cause (Fig. 10), or the locations of those working on COVID (Fig. S8B).

However, this also exemplifies how misleading missing data can be: Wikidata currently has highly inconsistent coverage of companies that are not publicly listed, which heavily biases the results. For example, the current lack of yearly updated socio-economic information such as *unemployment rates* [P1198] and *nominal GDP* [P2131] for countries in Wikidata limits the use of the knowledge graph for the study of the effect of the pandemic on global economies, although this is theoretically possible. Likewise, Wikidata is very incomplete with respect to COVID-19-related regulations like stay-at-home orders, school closures or policies regarding face masks. Standardised methods to audit and validate Wikidata's content on various topics are still under investigation and development [97].

4. Discussion

Many knowledge graphs have been rapidly developed to represent various types of COVID-19-related information, including government responses [22], epidemiology [108], clinical data [73], scholarly outputs and outcomes [46], economic impacts [8], physiopathology [24], social networking [23] among other features related to the

CORRECTED PROOF

COVID-19 pandemic. These semantic databases are mainly built using a combination of human efforts and crowdsourcing techniques [22]. Such resources can also be developed through the automatic extraction – using natural language processing techniques – of information from scholarly publications about the outbreak, as is the case with the COVID-19 Open Research Dataset [106].

Despite the importance of such resources, they tend to cover a narrow range of aspects of the disease, and despite the challenges (cf. Section 1.1), more integrated approaches are necessary to support advanced decision-making related to the outbreak. In response, integrated semantic databases have been launched to combine more divergent information, such as CIDO (combining clinical data with genomics) [37] and COVID-19 data hub (combining epidemiological data with social interactions) [35].

While clearly a valuable part of the data ecosystem, these projects rely on small groups of data curators, a model that has struggled to keep pace when data and scholarly literature grow sharply, as is the case with topics like COVID-19 [53]. This observation fits with the considerably limited volume of knowledge graphs exclusively enriched and verified by a dedicated expert group – such as OpenCyc – when compared to the volume of open and collaborative knowledge graphs, particularly Wikidata, YAGO, DBpedia and Freebase [30].

Whereas most knowledge graphs tend to be specialized and developed by a limited team, Wikidata deliberately takes a multidisciplinary, multilingual position anchored in the linked open data ecosystem. It is this breadth, combined with its interoperability, that makes it unique among even other user-generated collaborative projects. Indeed, it becomes uniquely suited to highly dynamic topics such as the COVID-19 pandemic [104,105]. In comparison to other resources like DBpedia, Wikidata is not just edited by machines and built from data automatically extracted from textual resources like Wikipedia [60]. Wikidata complements automated edits from trusted sources with enrichments and adjustments by a community of over 25000 active human users on a daily basis[66] and is released under the CC0 license allowing the free and unconditional reuse and interoperability of its information in other systems and datasets and consequently the growth of interest of many people in using, enriching and adjusting it [97]. By being highly multilingual, its human-readability extends well beyond English to support international contributions and reuse [97,98]. Also, its flexible editing policy and RDF structure permit the easy creation of new classes, properties and data models to rapidly support emerging data topics [97,98]. One of the features of Wikidata is also providing hundreds of exemplary SPARQL queries,[67] which even beginner users can immediately explore and easily modify, assisted with features like default prefixes, autosuggestions, autocomplete and straightforward conversion between Wikidata identifiers and natural language [65]. As a result, Wikidata users do not have to be SPARQL experts to arrive at results that are useful to them.

These factors have facilitated Wikidata's rapid growth since its creation in 2012 into an interdisciplinary network of >90 million items, richly interconnected by more than a billion statements [97,98]. In the context of the COVID-19 outbreak, Wikidata has proven its efficiency in representing multiple facets of the pandemic ranging from biomedical information to social impacts. This stands in marked contrast to other integrated semantic graphs that only combine two to three distinct features of the pandemic (e.g. CIDO [37], COVID-19 data hub [35], COVID-19 Living Data[68] [50] and Knowledge4COVID-19[69] [44]) as shown in the "data model" and "Visualizing facets of COVID-19 via SPARQL" sections. This large-scale information is supported in multiple languages as explained in the "language representation" section and is matched to its equivalents in other semantic databases as revealed by the "database alignment" section. Moreover, the semantic nature of the SPARQL query language has enabled in-depth analysis of the multifaceted, multidisciplinary COVID-19 information in Wikidata. This confirms previous findings about the importance of querying COVID-19 semantic resources such as CIDO [37] to compare clinical information with other types of COVID-19 information and consequently to generate new insights into or new perspectives on characteristics of the disease or the pandemic [16]. The primary advantage of applying SPARQL to extract and visualize COVID-19 information from a generalized knowledge graph such as Wikidata when compared to domain-specific knowledge graphs developed for the pandemic like CIDO [37] is the possibility of integration of

[66] https://www.wikidata.org/wiki/Special:Statistics.
[67] https://w.wiki/pGw.
[68] https://covid-nma.com/.
[69] https://devpost.com/software/covid-19-kg.

CORRECTED PROOF

outbreak data with non-COVID-19 information such as economic, industrial, climatic and social facts that can be used to generate summary information to explain the reasons behind the dynamics of the studied pandemic.

Despite the advantages of collaborative editing and free reuse of open knowledge graphs like Wikidata to support and enrich COVID-19 information, these two features have several drawbacks related to data quality and legal concerns. It is true that the use of fully open licenses (CC0 or Public domain) in centralized knowledge graphs removes all legal barriers to their reuse in other knowledge graphs or to drive knowledge-based systems and encourages the development of intelligent support for tasks related to COVID-19. However, application of CC0 on these databases causes them not to integrate information for semantic resources and datasets with only partially open licenses (e.g. CC BY and MIT), as these licenses require either the attribution of the source work to authors or the use of the same license to process the data [36,77]. This situation is analogous to the status of regular group O red blood cells as a universal donor but restricted recipient [63].

It is worth noting that crowd-sourced collaborative editing is often prone to the law of diminishing returns: the quality of human curation reaches a certain point, beyond which it is difficult to achieve additional major improvements. For instance, the quality of Wikidata relies partly on, e.g. automatically extracted infoboxes, which will only be verified and checked by editors some time in the future. However, research shows that the Wikidata community is already quite responsive to the needs of the database for all practical purposes. It is also worth remembering that machine-based systems are not immune to that effect neither [69]. Although collaborative editing contributed to the development of large-scale information about all aspects of the disease, there are currently still significant gaps and biases in the dataset that can lead to imprecise results if not interpreted with caution. For example, the COVID-19 outbreaks on cruise[70] and naval[71] ships are better covered in Wikipedia than in Wikidata (or most other online resources). Similarly, scholarly citations are not yet evenly covered, since systematic curation will require more scalable workflows. Although many of these gaps are rapidly being addressed and closed over time, errors of omission and bias are inevitable to some extent. Such deficiencies can only be detected and solved by applying algorithms that assess data completeness of items included in a given class within open knowledge graphs. Solutions involve cross-checking knowledge bases or subsets of the same knowledgebase [7,18], systematically exposing the content of Wikidata to many eyes through its reuse in Wikipedia and SPARQL-based tools such as Scholia and COVID dashboards [70,105], and using knowledge graph learning techniques to update items directly from textual databases like scholarly publications [110] and electronic health records [87]. Moreover, collaborative editing can cause several inaccuracies in the declaration of statements in open knowledge graphs disregarding the metadata standards of the knowledge bases [89]. These inconsistencies can persist particularly when the database and the largely growing scholarly literature about COVID-19 is managed by a limited number of administrators[72] and can consequently cause matters about the trustworthiness of the reuse of data [89]. However, critical problems related to structural deficiencies in defining statements or to the inclusion of mistaken data in open knowledge graphs seem to happen less frequently in Wikidata [30]. Greater consistency of structure and accuracy is partly due to the involvement of more contributors in Wikidata than in other open knowledge graphs [30]. But it also stems from importing data from other rapidly-updated and curated databases (mainly from the linked open data cloud [98]) and from verification by overlapping methods (e.g. ShEx schemas,[73] SPARQL-based logical constraints and bot edits [31,97]). The data validation infrastructure of Wikidata seems to be in accordance with the latest updates in knowledge graph evaluation and refinement techniques [75,109] and explains in part the reasons behind the robustness of the data model of COVID-19 information in this open knowledge graph.

CORRECTED PROOF

5. Conclusion

In this research paper, we demonstrate the ability of open and collaborative knowledge graphs such as Wikidata to represent and integrate multidisciplinary COVID-19 information and the use of SPARQL to generate summary

[70] https://en.wikipedia.org/wiki/COVID-19_pandemic_on_cruise_ships.

[71] https://en.wikipedia.org/wiki/COVID-19_pandemic_on_naval_ships.

[72] As of February 18, 2021, there are only 62 Wikidata administrators, as shown at https://www.wikidata.org/wiki/Special:Statistics.

[73] The validation schemas for COVID-19 information in Wikidata are currently available at https://www.wikidata.org/wiki/Wikidata:WikiProject_COVID-19/Data_models.

visualizations about the infectious disease, the underlying pathogen, the resulting pandemic and related topics. We have shown how the community-driven approach to editing without centralized coordination has contributed to the success of Wikidata in tackling emerging and rapidly changing phenomena, such as the pandemic. We have also discussed the disadvantages of collaborative editing for systematic knowledge representation, mainly the difficulty of ensuring sustainability for COVID-19 information in open knowledge graphs, the tricky validation of conflicting semantic data, the lack of coverage of several aspects of the analyzed pandemic, and the significant underrepresentation of advanced semantics for several types of Wikidata statements. Then, we described how the Wikimedia Community is currently trying to solve them through a series of advanced technical and organizational solutions. As an open semantic resource in the RDF format, Wikidata has become a hub for COVID-19 knowledge due to its alignment with major external resources and to its broad multidisciplinarity. The insertion of information in the Linked Open Data format provides the flexibility to integrate data from many facets of COVID-19 data with non-COVID-19 data. By its multilingual structure, these inputs are contributed to (and reused by) people all over the world, with different backgrounds. Effectively, the WikiProject COVID-19 has made COVID-19 knowledge more FAIR: Findable, Accessible, Interoperable and Reusable [104].

An important aspect of Wikidata's FAIRness is the Wikidata SPARQL query service (https://query.wikidata.org) [104]. More than an endpoint, the query service provides a visual interface to create queries, and makes it easier for beginners to customize queries. Additionally, community-contributed data visualization tools like Scholia provide human-friendly interfaces to surf the data [70]. As shown here, SPARQL visualizations are an entry point for deeper insights into COVID-19, including the integration of data from various fields and resources, and acquiring valuable implicit knowledge, both regarding the biomedical facets of this still new disease, as well as into the societal details of the pandemic.

As Wikidata is community-oriented and broadly themed, virtually any researcher can take advantage of its knowledge, and contribute to it. SPARQL queries can complement and enrich research publications, providing both an overview of domain-specific knowledge for original research, as well as serving as the base for systematic reviews or scientometric studies. Of note, SPARQL queries can be inserted into living publications, which can keep up to date with the advancements both in human knowledge and its coverage on Wikidata.

Another part of FAIRness is user-friendly programmatic data access. Wikidata database dumps are available for download and local processing in RDF, JSON and XML formats (https://www.wikidata.org/wiki/Wikidata:Database_download). Beyond dumps, the Wikibase API makes data retrievable via HTTP requests, which facilitates integration into analysis and reuse workflows. API wrappers are also available for popular programming languages like R (https://cran.r-project.org/web/packages/WikidataR/) and Python (https://pypi.org/project/Wikidata/), further boosting interoperability.

Even though Wikidata is rich in COVID-19 knowledge, there is always room for improvement. As a collaborative endeavour, Wikidata and the WikiProject COVID-19 are likely to become further enriched over time. By the collective efforts of contributors, we hope that the database will grow in quality and coverage, supporting other types of information – such as the outcomes of the ongoing COVID-19-related research efforts – and contributing to higher pandemic preparedness globally.

CORRECTED PROOF

Conflict of interest

All authors of this paper are active members of WikiProject Medicine, the community curating clinical knowledge in Wikidata, and of WikiProject COVID-19, the community developing multidisciplinary COVID-19 information in Wikidata. DJ is a non-paid voluntary member of the Board of Trustees of the Wikimedia Foundation, the non-profit publisher of Wikipedia and Wikidata.

Data availability

Source files for most of the tables featured in this work are available at https://github.com/csisc/WikidataCOVID19SPARQL under the CC0 license [96]. Figures involved in this research study are available at https://w.wiki/

bnW mostly under the CC BY 4.0 License and their source SPARQL queries[74] are linked in the figure legends. Internet archive links for the cited URLs are made available at https://doi.org/10.5281/zenodo.4022591 thanks to ArchiveNow [6].

Supplementary data

Supplementary material containing Supplementary Tables S1-S8 and Supplementary Figures S1-10 is available at: http://dx.doi.org/10.3233/SW-210444.

Acknowledgements

The work done by Houcemeddine Turki, Mohamed Ali Hadj Taieb and Mohamed Ben Aouicha was supported by the Ministry of Higher Education and Scientific Research in Tunisia (MoHESR) in the framework of Federated Research Project PRFCOV19-D1-P1. The work done by Jose Emilio Labra Gayo was partially funded by the Spanish Ministry of Economy and Competitiveness (Society challenges: TIN2017-88877-R). The work done by Daniel Mietchen was supported in part by the Alfred P. Sloan Foundation under grant number G-2019-11458. The work done by Dariusz Jemielniak was funded by Polish National Science Center grant no 2019/35/B/HS6/01056. We thank the Wikidata community, Egon Willighagen (Maastricht University, Netherlands), Toby Hudson (University of Sydney, Australia), Adam Shorland (Wikimedia Deutschland, Germany), and Mahir Morshed (University of Illinois at Urbana-Champaign, United States of America) for useful comments and discussions about the topic of this research paper.

References

[1] D. Abián, F. Guerra, J. Martínez-Romanos and R. Trillo-Lado, Wikidata and DBpedia: A comparative study, in: *Semanitic Keyword-Based Search on Structured Data Sources*, Springer, 2017, pp. 142–154. doi:10.1007/978-3-319-74497-1_14.

[2] Addshore, D. Mietchen and E. Willighagen, Wikidata queries around the SARS-CoV-2 virus and pandemic, *Zenodo* (2020). doi:10.5281/zenodo.3977414.

[3] S. Allison-Cassin and D. Scott, Wikidata: A platform for your library's linked open data, *Code4Lib Journal* **40** (2018).

[4] D. Alocci, J. Mariethoz, O. Horlacher, J.T. Bolleman, M.P. Campbell and F. Lisacek, Property graph vs RDF triple store: A comparison on glycan substructure search, *PloS one* **10**(12) (2015), e0144578. doi:10.1371/journal.pone.0144578.

[5] R. Angles, H. Thakkar and D. Tomaszuk, RDF and property graphs interoperability: Status and issues, in: *Proceedings of the 13th Alberto Mendelzon International Workshop on Foundations of Data Management*, CEUR-WS.org, 2019, Paper 1.

[6] M. Aturban, M. Kelly, S. Alam, J.A. Berlin, M.L. Nelson and M.C. Weigle, ArchiveNow: Simplified, extensible, multi-archive preservation, in: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, ACM, 2018, pp. 321–322. doi:10.1145/3197026.3203880.

[7] V. Balaraman, S. Razniewski and W. Nutt, Recoin: Relative completeness in Wikidata, in: *Companion Proceedings of the Web Conference 2018*, ACM, 2018, pp. 1787–1792. doi:10.1145/3184558.3191641.

[8] L. Bellomarini, M. Benedetti, A. Gentili, R. Laurendi, D. Magnanimi, A. Muci and E. Sallinger, 2020, COVID-19 and Company Knowledge Graphs: Assessing Golden Powers and Economic Impact of Selective Lockdown via AI Reasoning. arXiv preprint arXiv:2004.10119.

[9] O. Benjakob, R. Aviram and J.A. Sobel, Meta-Research: Citation needed? Wikipedia and the COVID-19 pandemic, *bioRxiv* (2021). doi:10.1101/2021.03.01.433379.

[10] A. Boccone and R. Rivelli, The bibliographic metadata in Wikidata: Wikicite and the «Bibliothecae.it»case study, *Bibliothecae.it* **8.1** (2019), 227–248. doi:10.6092/issn.2283-9364/9503.

[11] S. Burgstaller-Muehlbacher, A. Waagmeester, E. Mitraka, J. Turner, T. Putman, J. Leong et al., Wikidata as a semantic framework for the gene wiki initiative, *Database* **2016** (2016), baw015. doi:10.1093/database/baw015.

[12] L. Chaari and O. Golubnitschaja, Covid-19 pandemic by the "real-time" monitoring: The Tunisian case and lessons for global epidemics in the context of 3PM strategies, *EPMA journal* **11**(2) (2020), 133–138. doi:10.1007/s13167-020-00207-0.

[74] The source code of the SPARQL queries used in this work are also made available at https://web.archive.org/web/20200914223401/https://www.wikidata.org/wiki/Wikidata:WikiProject_COVID-19/Queries/SPARQL_Study.

CORRECTED PROOF

[13] J. Chrzanowski, J. Sołek and D. Jemielniak, Assessing public interest based on Wikipedia's most visited medical articles during the SARS-CoV-2 outbreak: Search trends analysis, *Journal of medical Internet research* **23**(4) (2021), e26331. doi:10.2196/26331.
[14] L. Ciechanowski, D. Jemielniak and P.A. Gloor, TUTORIAL: AI research without coding: The art of fighting without fighting: Data science for qualitative researchers, *Journal of Business Research* **117** (2020), 322–330. doi:10.1016/j.jbusres.2020.06.012.
[15] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum, Hillsdale, 1988. ISBN: 978-0-8058-0283-2.
[16] A.S. Dadzie and M. Rowe, Approaches to visualising linked data: A survey, *Semantic Web* **2**(2) (2011), 89–124. doi:10.3233/SW-2011-0037.
[17] F. Darari, COVIWD: COVID-19 Wikidata dashboard, *Jurnal Ilmu Komputer dan Informasi* **14**(1) (2021), 39–47. doi:10.21609/jiki.v14i1.941.
[18] F. Darari, S. Razniewski, R.E. Prasojo and W. Nutt, Enabling fine-grained RDF data completeness assessment, in: *International Conference on Web Engineering*, Springer, Cham, 2016, pp. 170–187. doi:10.1007/978-3-319-38791-8_10.
[19] G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi and R. Rosati, Using ontologies for semantic data integration, in: *A Comprehensive Guide Through the Italian Database Research over the Last 25 Years*, Springer, 2018, pp. 187–202. doi:10.1007/978-3-319-61893-7_11.
[20] G. De Melo and G. Weikum, Towards universal multilingual knowledge bases, in: *Principles, Construction, and Applications of Multilingual Wordnets*, Proceedings of the 5th Global WordNet Conference (GWC 2010), Narosa Publishing, New Delhi, India, 2010, pp. 149–156.
[21] J. Debattista, C. Lange, S. Auer and D. Cortis, Evaluating the quality of the LOD cloud: An empirical investigation, *Semantic Web* **9**(6) (2018), 859–901. doi:10.3233/SW-180306.
[22] A. Desvars-Larrive, E. Dervic, N. Haug, T. Niederkrotenthaler, J. Chen, A. Di Natale et al., A structured open dataset of government interventions in response to COVID-19, *Scientific data* **7**(1) (2020), 285. doi:10.1038/s41597-020-00609-9.
[23] D. Dimitrov, E. Baran, P. Fafalios, R. Yu, X. Zhu, M. Zloch and S. Dietze, TweetsCOV19–A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic, 2020, arXiv preprint arXiv:2006.14492.
[24] D. Domingo-Fernández, S. Baksi, B. Schultz, Y. Gadiya, R. Karki, T. Raschka et al., COVID-19 knowledge graph: A computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology, *BioRxiv* (2020). doi:10.1101/2020.04.14.040667.
[25] S. Dubey, P. Biswas, R. Ghosh, S. Chatterjee, M.J. Dubey, S. Chatterjee et al., Psychosocial impact of COVID-19, *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* **14**(5) (2020), 779–788. doi:10.1016/j.dsx.2020.05.035.
[26] D.M. Eberhand, G.F. Simons and C.D. Fennig, *Ethnologue: Languages of the World*, SIL International, Dallas, Texas, 2020.
[27] L. Ehrlinger and W. Wöß, Towards a definition of knowledge graphs, *CEUR Workshop Proceedings* **1695** (2016), 1–4.
[28] S. Ekins, D. Mietchen, M. Coffee, T.P. Stratton, J.S. Freundlich, L. Freitas-Junior et al., Open drug discovery for the Zika virus. *F1000Research*, 2016 (2016), p. 5:150. doi:10.12688/f1000research.8013.1.
[29] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez and D. Vrandečić, Introducing Wikidata to the Linked Data Web, in: *International Semantic Web Conference*, Springer, Cham, 2014, pp. 50–65. doi:10.1007/978-3-319-11964-9_4.
[30] M. Färber, F. Bartscherer, C. Menne and A. Rettinger, Linked data quality of DBpedia, *Freebase, OpenCyc, Wikidata, and YAGO. Semantic Web* **9**(1) (2018), 77–129. doi:10.3233/SW-170275.
[31] M. Farda-Sarbas, H. Zhu, M.F. Nest and C. Müller-Birn, Approving automation: Analyzing requests for permissions of bots in Wikidata, in: *Proceedings of the 15th International Symposium on Open Collaboration*, 2019, pp. 1–10. doi:10.1145/3306446.3340833.
[32] R.D. Finn, P.P. Gardner and A. Bateman, Making your database available through Wikipedia: The pros and cons, *Nucleic acids research* **40.D1** (2012), D9–D12. doi:10.1093/nar/gkr1195.
[33] P. Fontelo, F. Liu, S. Leon, A. Abrahamane and M. Ackerman (PICO Linguist), BabelMeSH: Development and partial evaluation of evidence-based multilanguage search tools for MEDLINE/PubMed, in: *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics*, IOS Press, 2007, p. 817.
[34] F. Freitas, S. Schulz and E. Moraes, Survey of current terminologies and ontologies in biology and medicine, *Reciis* **3**(1) (2009), 7–18. doi:10.3395/reciis.v3i1.239en.
[35] E. Guidotti and D. Ardia, COVID-19 data hub, *Journal of Open Source Software* **5**(51) (2020), 2376. doi:10.21105/joss.02376.
[36] G. Hagedorn, D. Mietchen, R.A. Morris, D. Agosti, L. Penev, W.G. Berendsohn and D. Hobern, Creative commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information, *ZooKeys* **150** (2011), 127. doi:10.3897/zookeys.150.2189.
[37] Y. He, H. Yu, E. Ong, Y. Wang, Y. Liu, A. Huffman et al., CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis, *Scientific data* **7**(1) (2020), 181. doi:10.1038/s41597-020-0523-6.
[38] J.M. Heilman and A.G. West, Wikipedia and medicine: Quantifying readership, editors, and the significance of natural language, *Journal of Medical Internet Research* **17**(3) (2015), e62. doi:10.2196/jmir.4069.
[39] A. Henriksson, M. Skeppstedt, M. Kvist, M. Duneld and M. Conway, Corpus-driven terminology development: Populating Swedish SNOMED CT with synonyms extracted from electronic health records, in: *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, 2013, pp. 36–44.
[40] D. Hernández, A. Hogan and M. Krötzsch, Reifying RDF: What works well with Wikidata, in: *Proceedings of the Scalable Semantic Web Systems Workshop (SSWS 2015) at International Semantic Web Conference*, CEUR-WS.org, 2015, pp. 32–47.
[41] P. Hitzler and K. Janowicz, Linked data, big data, and the 4th paradigm, *Semantic Web* **4**(3) (2013), 233–235. doi:10.3233/SW-130117.
[42] X. Hu, R. Rousseau and J. Chen, On the definition of forward and backward citation generations, *Journal of Informetrics* **5**(1) (2011), 27–36. doi:10.1016/j.joi.2010.07.004.
[43] J.W. Huss III., C. Orozco, J. Goodale, C. Wu, S. Batalov, T.J. Vickers et al., A gene wiki for community annotation of gene function, *PLoS Biol* **6**(7) (2008), e175.

CORRECTED PROOF

[44] E. Iglesias, S. Jozashoori, D. Chaves-Fraga, D. Collarana and M.E. Vidal, SDM-RDFizer: An RML Interpreter for the Efficient Creation of RDF Knowledge Graphs, 2020, preprint arXiv:2008.07176.
[45] A. Ismayilov, D. Kontokostas, S. Auer, J. Lehmann and S. Hellmann, Wikidata through the eyes of DBpedia, *Semantic Web* **9**(4) (2018), 493–503. doi:10.3233/SW-170277.
[46] M.Y. Jaradeh, A. Oelen, K.E. Farfar, M. Prinz, J. D'Souza, G. Kismihók et al., Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge, in: *Proceedings of the 10th International Conference on Knowledge Capture*, 2019, pp. 243–246. doi:10.1145/3360901.3364435.
[47] D. Jemielniak, *Common Knowledge?: An Ethnography of Wikipedia*, Stanford University Press, Stanford, 2014. ISBN: 978-0804789448.
[48] D. Jemielniak, G. Masukume and M. Wilamowski, The most influential medical journals according to Wikipedia: Quantitative analysis, *Journal of Medical Internet Research* **21**(1) (2019), e11429. doi:10.2196/11429.
[49] N. Jetté, H. Quan, B. Hemmelgarn, S. Drosler, C.D.-G. Maass et al., The development, evolution, and modifications of ICD-10: Challenges to the international comparability of morbidity data, *Medical Care* **48**(12) (2010), 1105–1110. doi:10.1097/MLR.0b013e3181ef9d3e.
[50] S. Juul, N. Nielsen, P. Bentzer, A.A. Veroniki, L. Thabane, A. Linder et al., Interventions for treatment of COVID-19: A protocol for a living systematic review with network meta-analysis including individual patient data (The LIVING Project), *Systematic Reviews* **9** (2020), 108. doi:10.1186/s13643-020-01371-0.
[51] L.-A. Kaffee and E. Simperl, Analysis of editors' languages in Wikidata, in: *Proceedings of the 14th International Symposium on Open Collaboration*, ACM, 2018, p. 21. doi:10.1145/3233391.3233965.
[52] L.A. Kaffee, A. Piscopo, P. Vougiouklis, E. Simperl, L. Carr and L. Pintscher, A glimpse into babel: An analysis of multilinguality in Wikidata, *Proceedings of the 13th International Symposium on Open Collaboration* (2017), 14. doi:10.1145/3125433.3125465.
[53] D. Kagan, J. Moran-Gilad and M. Fire, Scientometric trends for coronaviruses and other emerging viral infections, *GigaScience* **9**(8) (2020), giaa085. doi:10.1093/gigascience/giaa085.
[54] B. Keegan, D. Gergle and N. Contractor, Hot off the wiki: Structures and dynamics of Wikipedia's coverage of breaking news events, *American behavioral scientist* **57**(5) (2013), 595–622. doi:10.1177/0002764212469367.
[55] B.C. Keegan and J.R. Brubaker, 'Is' to 'was' coordination and commemoration in posthumous activity on Wikipedia biographies, in: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ACM, 2015, pp. 533–546. doi:10.1145/2675133.2675238.
[56] B.C. Keegan and C. Tan, A Quantitative Portrait of Wikipedia's High-Tempo Collaborations during the 2020 Coronavirus Pandemic, 2020, arXiv preprint arXiv:2006.08899.
[57] A. Kirk, Data visualisation: A handbook for data driven design, Sage, 2016, ISBN:9781473966314.
[58] M. Klein and A. Kyrios, VIAFbot and the integration of library data on Wikipedia, *Code4lib journal* (2013), 22.
[59] P. Konieczny, Adhocratic governance in the Internet age: A case of Wikipedia, *Journal of Information Technology & Politics* **7**(4) (2010), 263–283. doi:10.1080/19331681.2010.489408.
[60] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes et al., DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia, *Semantic Web* **6**(2) (2015), 167–195. doi:10.3233/SW-140134.
[61] F. Liu, P. Fontelo and M. Ackerman, BabelMeSH: Development of a cross-language tool for MEDLINE/PubMed, in: *AMIA Annual Symposium Proceedings, American Medical Informatics Association*, 2006, p. 1012.
[62] Y. Liu, W.K.B. Chan, Z. Wang, J. Hur, J. Xie, H. Yu and Y. He, Ontological and bioinformatic analysis of anti-coronavirus drugs and their Implication for drug repurposing against COVID-19, *Preprints* (2020), 2020030413. doi:10.20944/preprints202003.0413.v1.
[63] D.M. Lublin, Universal RBCs, *Transfusion* **40**(11) (2000), 1285–1289. doi:10.1046/j.1537-2995.2000.40111285.x.
[64] M.S. Majumder and K.D. Mandl, Early in the epidemic: Impact of preprints on global discourse about COVID-19 transmissibility, *The Lancet Global Health* **8**(5) (2020), e627–e630. doi:10.1016/S2214-109X(20)30113-3.
[65] S. Malyshev, M. Krötzsch, L. González, J. Gonsior and A. Bielefeldt, Getting the most out of Wikidata: Semantic technology usage in Wikipedia's knowledge graph, in: *International Semantic Web Conference*, Springer, Cham, 2018, pp. 376–394. doi:10.1007/978-3-030-00668-6_23.
[66] M. Manske, U. Böhme, C. Püthe and M. Berriman, GeneDB and Wikidata, *Wellcome open research* **4** (2019), 114. doi:10.12688/wellcomeopenres.15355.2.
[67] D. Mietchen, State of WikiCite in 2020, in: *Workshop on Open Citations and Open Scholarly Metadata 2020*, 2020. doi:10.5281/zenodo.4019954.
[68] D. Mietchen, G. Hagedorn, E. Willighagen, M. Rico, A. Gómez-Pérez, E. Aibar, K. Rafes, C. Germain, A. Dunning, L. Pintscher and D. Kinzler, Enabling open science: Wikidata for research (Wiki4R), *Research Ideas and Outcomes* **1** (2015), e7573. doi:10.3897/rio.1.e7573.
[69] J.M. Mortensen, E.P. Minty, M. Januszyk, T.E. Sweeney, A.L. Rector, N.F. Noy and M.A. Musen, Using the wisdom of the crowds to find critical errors in biomedical ontologies: A study of SNOMED CT, *Journal of the American Medical Informatics Association* **22**(3) (2015), 640–648. doi:10.1136/amiajnl-2014-002901.
[70] F.Å. Nielsen, D. Mietchen and E. Willighagen, Scholia, scientometrics and Wikidata, in: *European Semantic Web Conference*, Springer, Cham, 2017, pp. 237–259. doi:10.1007/978-3-319-70407-4_36.
[71] J. Nunn, T. Shafee, S. Chang, R. Stephens, J. Elliott, S. Oliver et al., *Standardised Data on Initiatives-STARDIT: Alpha Version. OSF Preprints*, 2019. doi:10.31219/osf.io/5q47h.
[72] C. Ordun, S. Purushotham and E. Raff, Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs, 2020, preprint arXiv:2005.03082.

CORRECTED PROOF

[73] M. Ostaszewski, A. Mazein, M.E. Gillespie, I. Kuperstein, A. Niarakis, H. Hermjakob et al., COVID-19 disease map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms, *Scientific data* **7**(1) (2020), 136. doi:10.1038/s41597-020-0477-8.
[74] R. Page, Towards a biodiversity knowledge graph, *Research Ideas and Outcomes* **2** (2016), e8767. doi:10.3897/rio.2.e8767.
[75] H. Paulheim, Knowledge graph refinement: A survey of approaches and evaluation methods, *Semantic Web* **8**(3) (2017), 489–508. doi:10.3233/SW-160218.
[76] T. Pellissier Tanon, D. Vrandečić, S. Schaffert, T. Steiner and L. Pintscher, From Freebase to Wikidata: The great migration, in: *Proceedings of the 25th International Conference on World Wide Web*, ACM, 2016, pp. 1419–1428. doi:10.1145/2872427.2874809.
[77] L. Penev, D. Mietchen, V. Chavan, G. Hagedorn, V. Smith, D. Shotton et al., Strategies and guidelines for scholarly publishing of biodiversity data, *Research Ideas and Outcomes* **3** (2017), e12431. doi:10.3897/rio.3.e12431.
[78] E. Prud'hommeaux, J.E. Labra Gayo and H.S. Solbrig, Shape Expressions: An RDF validation and transformation language, in: *Proceedings of the 10th International Conference on Semantic Systems*, ACM, 2014, pp. 32–40. doi:10.1145/2660517.2660523.
[79] T.E. Putnam, S. Lelong, S. Burgstaller-Muehlbacher, A. Waagmeester, C. Diesh, N. Dunn et al., WikiGenomes: An open web application for community consumption and curation of gene annotation data in Wikidata, *Database* (2017), 2017, bax025. doi:10.1093/database/bax025.
[80] N. Queralt-Rosinach, G.S. Stupp, T.S. Li, M. Mayers, M.E. Hoatlin, M. Might et al., Structured reviews for data and knowledge-driven research, *Database* **2020** (2020), baaa015. doi:10.1093/database/baaa015.
[81] A. Rahimi, T. Baldwin and K. Verspoor, WikiUMLS: Aligning UMLS to Wikipedia via Cross-lingual Neural Ranking, 2020, arXiv preprint arXiv:2005.01281.
[82] RDA COVID-19 Working Group, *Recommendations and Guidelines on Data Sharing. Research Data Alliance*, 2020. doi:10.15497/rda00052.
[83] M. Rico, N. Mihindukulasooriya, D. Kontokostas, H. Paulheim, S. Hellmann and A. Gómez-Pérez, Predicting incorrect mappings: A data-driven approach applied to DBpedia, in: *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, ACM, 2018, pp. 323–330. doi:10.1145/3167132.3167164.
[84] D. Ringler and H. Paulheim, One knowledge graph to rule them all? Analyzing the differences between DBpedia, YAGO, Wikidata & co, in: *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, Springer, 2017, pp. 366–372. doi:10.1007/978-3-319-67190-1_33.
[85] P. Ristoski, G.K.D. De Vries and H. Paulheim, A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web, in: *International Semantic Web Conference*, Springer, Cham, 2016, pp. 186–194. doi:10.1007/978-3-319-46547-0_20.
[86] R.C. Rodgers, Z. Sherwin, H. Lamberts and I.M. Okkes, ICPC multilingual collaboratory: A web-and unicode-based system for distributed editing/translating/viewing of the multilingual international classification of primary care, in: *Medinfo 2004*, IOS Press, 2004, pp. 425–429. doi:10.3233/978-1-60750-949-3-425.
[87] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng and D. Sontag, Learning a health knowledge graph from electronic medical records, *Scientific reports* **7**(1) (2017), 1–11. doi:10.1038/s41598-017-05778-z.
[88] J. Samuel, Towards understanding and improving multilingual collaborative ontology development in Wikidata, in: *Companion of the Web Conference 2018 on the Web Conference*, ACM, 2018, pp. 23–27.
[89] L.M. Schriml, M. Chuvochina, N. Davies, E.A. Eloe-Fadrosh, R.D. Finn, P. Hugenholtz et al., COVID-19 pandemic reveals the peril of ignoring metadata standards, *Scientific data* **7**(1) (2020), 188. doi:10.1038/s41597-020-0524-5.
[90] H. Sebei, M.A. Hadj Taieb and M. Ben Aouicha, Review of social media analytics process and big data pipeline, *Social Network Analysis and Mining* **8**(1) (2018), 30. doi:10.1007/s13278-018-0507-0.
[91] T. Shafee, G. Masukume, L. Kipersztok, D. Das, M. Häggström and J. Heilman, Evolution of Wikipedia's medical content: Past, present and future, *J Epidemiol Community Health* **71**(11) (2017), 1122–1129. doi:10.1136/jech-2016-208601.
[92] D. Taraborelli, L. Pintscher, D. Mietchen and S. Rodlund, WikiCite 2017 report, *Figshare* (2017). doi:10.6084/m9.figshare.5648233.
[93] A.R. Terryn, V. Hoste and E. Lefever, In no uncertain terms: A dataset for monolingual and multilingual automatic term extraction from comparable corpora, *Language Resources and Evaluation* **54** (2019), 385–418. doi:10.1007/s10579-019-09453-9.
[94] I. Tiddi, M. d'Aquin and E. Motta, Walking linked data: A graph traversal approach to explain clusters, in: *Proceedings of the 5th International Conference on Consuming Linked Data-Volume 1264*, ACM, 2014, pp. 73–84. doi:10.5555/2877789.2877796.
[95] H. Turki, M.A. Hadj Taieb and M. Ben Aouicha, 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** (2018), 217–218. doi:10.1016/j.jbi.2018.05.011.
[96] H. Turki, M.A. Hadj Taieb, T. Shafee, T. Lubiana, D. Jemielniak, M. Ben Aouicha, J.E. Labra Gayo, M. Banat, D. Das and D. Mietchen, Csisc/WikidataCOVID19SPARQL: Data about Wikidata coverage of Covid-19, *Zenodo* (2020). doi:10.5281/zenodo.4022591.
[97] H. Turki, D. Jemielniak, M.A. Hadj Taieb, J.E. Labra Gayo, M. Ben Aouicha, M. Banat, T. Shafee, E. Prud'Hommeaux, T. Lubiana, D. Das and D. Mietchen, Using logical constraints to validate information in collaborative knowledge graphs: A study of COVID-19 on Wikidata, *Zenodo* (2020). doi:10.5281/zenodo.4008358.
[98] H. Turki, T. Shafee, M.A. Hadj Taieb, M. Ben Aouicha, D. Vrandečić, D. Das and H.H. Wikidata, A large-scale collaborative ontological medical database, *Journal of Biomedical Informatics* **99** (2019), 103292. doi:10.1016/j.jbi.2019.103292.
[99] United Nations Development Programme. Human Development Report 2020 The Next Frontier: Human Development and the Anthropocene. United Nations Development Programme, 2020, pp. 343–346. ISBN 978-92-1-126442-5.
[100] N. Vincent, I. Johnson and B. Hecht, Examining Wikipedia with a broader lens: Quantifying the value of Wikipedia's relationships with other large-scale online communities, in: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ACM, 2018, pp. 1–13. doi:10.1145/3173574.3174140.

CORRECTED PROOF

[101] P. Vossen, R. Agerri, I. Aldabe, A. Cybulska, M. van Erp, A. Fokkens et al., Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news, *Knowledge-Based Systems* **110** (2016), 60–85. doi:10.1016/j.knosys.2016.07.013.

[102] D. Vrandečić and M. Krötzsch, Wikidata: A free collaborative knowledgebase, *Communications of the ACM* **57**(10) (2014), 78–85. doi:10.1145/2629489.

[103] A. Waagmeester, L. Schriml and A.I. Su, Wikidata as a linked-data hub for biodiversity data, *Biodiversity Information Science and Standards* **3** (2019), e35206. doi:10.3897/biss.3.35206.

[104] A. Waagmeester, G. Stupp, S. Burgstaller-Muehlbacher, B.M. Good, G. Malachi, O.L. Griffith et al., Wikidata as a knowledge graph for the life sciences, *eLife* **9** (2020), e52614. doi:10.7554/eLife.52614.

[105] A. Waagmeester, E.L. Willighagen, A.I. Su, M. Kutmon, J.E. Labra Gayo, D. Fernández-Álvarez et al., A protocol for adding knowledge to Wikidata: Aligning resources on human coronaviruses, *BMC Biology* **19** (2021), 12. doi:10.1186/s12915-020-00940-y.

[106] L.L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide et al., 2020, CORD-19: The Covid-19 Open Research Dataset. arXiv preprint, https://www.ncbi.nlm.nih.gov/pmc/articles/pmc7251955/ arXiv:2004.10706.

[107] L. Wyatt, P. Ayers, M. Proffitt, D. Mietchen, E. Seiver, A. Stinson, D. Taraborelli, C. Virtue, J. Tud and J. Curiel, WikiCite Annual Report, 2019–20. *Zenodo* (2020). doi:10.5281/zenodo.3869809.

[108] B. Xu and M.U. Kraemer (Data Curation Group), Open access epidemiological data from the COVID-19 outbreak, *The Lancet Infectious Diseases* **20**(5) (2020), 534. doi:10.1016/S1473-3099(20)30119-5.

[109] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann and S. Auer, Quality assessment for linked data: A survey, *Semantic Web* **7**(1) (2016), 63–93. doi:10.3233/SW-150175.

[110] Y. Zhang, H. Lin, Z. Yang, J. Wang, S. Zhang, Y. Sun and L. Yang, A hybrid model based on neural networks for biomedical relation extraction, *Journal of biomedical informatics* **81** (2018), 83–92. doi:10.1016/j.jbi.2018.03.011.

[111] Z. Zhang, W. Yao, Y. Wang, C. Long and X. Fu, Wuhan and Hubei COVID-19 mortality analysis reveals the critical role of timely supply of medical resources, *The Journal of infection* **81**(1) (2020), 147. doi:10.1016/j.jinf.2020.03.018.

CORRECTED PROOF