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

Editor(s): Armin Haller, Australian National University, Australia
Solicited review(s): Pouya Ghiasnezhad Omran, Australian National University, Australia; Gengchen Mai, University of California, Santa Barbara, United States of America

Houcemeddine Turki[a] (ORCID: 0000-0003-3492-2014), Mohamed Ali Hadj Taieb[b] (ORCID: 0000-0002-2786-8913), Thomas Shafee[c] (ORCID: 0000-0002-2298-7593), Tiago Lubiana[d] (ORCID: 0000-0003-2473-2313), Dariusz Jemielniak[e] (ORCID: 0000-0002-3745-7931), Mohamed Ben Aouicha[b] (ORCID: 0000-0002-2277-5814), Jose Emilio Labra Gayo[f] (ORCID: 0000-0001-8907-5348), Eric A. Youngstrom[g] (ORCID: 0000-0003-2251-6860), Mus'ab Banat[h] (ORCID: 0000-0001-9132-3849), Diptanshu Das[i] (ORCID: 0000-0002-7221-5022), Daniel Mietchen[j,1*] (ORCID: 0000-0001-9488-1870), on behalf of WikiProject COVID-19[2]

[a] *Faculty of Medicine of Sfax, University of Sfax, Sfax, Tunisia*
[b] *Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia*
[c] *La Trobe University, Melbourne, Victoria, Australia*
[d] *Computational Systems Biology Laboratory, University of São Paulo, São Paulo, Brazil*
[e] *Department of Management in Networked and Digital Societies, Kozminski University, Warsaw, Poland*
[f] *Web Semantics Oviedo (WESO) Research Group, University of Oviedo, Spain*
[g] *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*
[h] *Faculty of Medicine, Hashemite University, Zarqa, Jordan*
[i] *Institute of Child Health (ICH), Kolkata, India*
[i] *Medica Superspecialty Hospital, Kolkata, India*
[j] *School of Data Science, University of Virginia, Charlottesville, Virginia, United States of America*

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. A web-scale platform for broader computer-supported cooperative work and linked open data, it can be queried in multiple ways in near real time by specialists, automated tools and the public, including via SPARQL, a semantic query 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

[1*] Corresponding author. E-mail: dm7gn@virginia.edu.

[2] Project Member: 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.

1. Introduction

The COVID-19 pandemic is complex and multifaceted and touches on almost every aspect of current life [1]. 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 [2]. There are already attempts at creating community-based ontologies of COVID-19 knowledge and data [3], as well as efforts to aggregate expert data [4]. Many open data initiatives have been started spontaneously [5-7]. The interconnected, multidisciplinary, and international nature of the pandemic creates both challenges and opportunities for using knowledge graphs [2-5, 8-12].

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 [2-5, 8-12]. While the manuscript as a whole emphasizes the opportunities, we think it is worthwhile to highlight some of the challenges early on.

1.1. Data integration challenges

The integration of different data sources always poses a range of challenges [13], 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).

With respect to integrating COVID-19 data in particular, a number of further challenges need to be considered. We will refer to them collectively as COVID-19 data challenges, of which we will briefly outline four major ones: First, human knowledge about the COVID-19 disease, the underlying pathogen and the resulting pandemic is evolving rapidly [14], 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 [15]. 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 [1], the curated data about it should ideally be easily and reliably accessible for humans and machines across a broad range of use cases [16].

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 [17-19], 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.

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 a snapshot of how the COVID-19 knowledge graph of Wikidata can be used to support computer applications, particularly the SPARQL-based visualization of 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 [20].

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 [21]. 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 [22]: 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[3].

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 [20, 23], and its existing community has been using it to capture COVID-19-related knowledge right from the start. 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 [15]. Third, both the Wikidata platform and the Wikidata community are highly multifaceted, multilingual and multidisciplinary [24, 25]. 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 [20], 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 [26] or through overlap with the Wikipedia community that has been covering disasters for two decades.[4]

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 [27]) or multilinguality (e.g. how to handle concepts that are hard to translate [28]), 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 [29].

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 [20]. 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 [30]. 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[5], a semantic query language dynamically

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

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

[5] 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/.

extracting triple information from large-scale knowledge graphs.

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.

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 Wikidata as a collaborative multidisciplinary large-scale knowledge graph can host COVID-19 data, integrate it with non-COVID-19 information and feed computer applications in an open and transparent way.

2.1. Data model

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 [19]. 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 [19].

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 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 are also organised as RDF triples [23]. This process is called reification, and it is a common feature of many knowledge graphs such as DBpedia, Freebase, and YAGO [17]. Although DBpedia and Freebase apply reification in a similar setting as in Wikidata, YAGO chooses to use N-Quads to represent the characteristics of a statement, implying that the additional feature is linked to the statement as a couple without the use of any predicate [17].

This comes together to create an integrated network of over 90 million items interlinked by over a billion statements. Its volume, variety, velocity and veracity place it well in the scope of ‘big data’ approaches [31, 32]. The advantage of RDF over other competing semantic data formats, particularly *property graph*, is that it applies reference schemas and consistency rules before assigning predicates to statements [33].

Entries in RDF triple stores are predefined entities, rather than simple text strings, and structured into uni-directional statements [34]. 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 graph* [33], 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 [34].

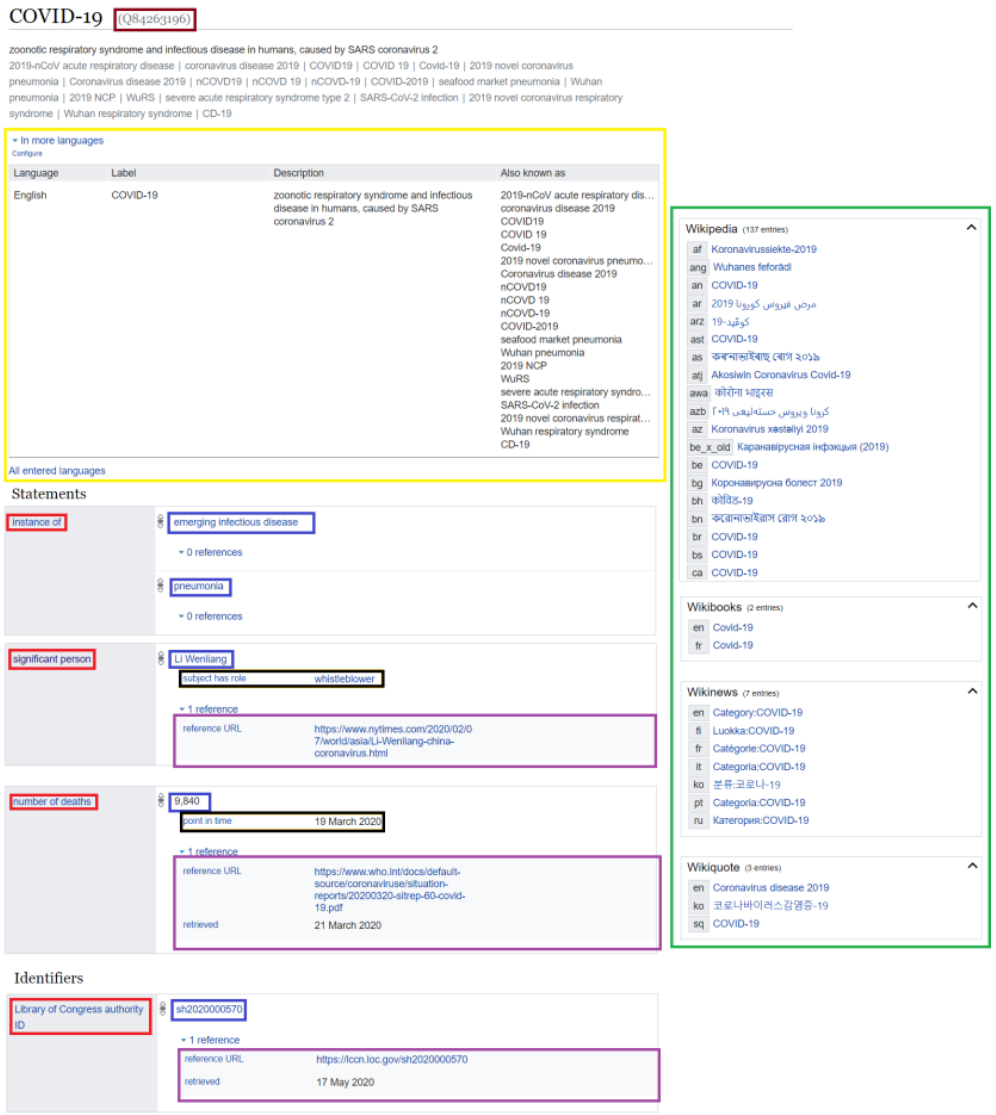

Fig. 1. Data Structure of a Wikidata item. The simple, consistent structure of a Wikidata item makes it both human- and machine-readable. Each Wikidata item has a unique identifier (Brown). Items can have labels, descriptions and aliases in multiple languages (Yellow). They can include any number of 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/auE. License: CC-BY-SA-4.0.

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 [26] and led to the formation of WikiProject Zika Corpus[6]. This Zika project—itself inspired by dedicated Wikiprojects for Medicine[7] and for Source Metadata[8]—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

6 https://www.wikidata.org/wiki/Wikidata:WikiProject_Zika_Corpus

7 https://www.wikidata.org/wiki/Wikidata:WikiProject_Medicine

8 https://www.wikidata.org/wiki/Wikidata:WikiProject_Source_MetaData

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.

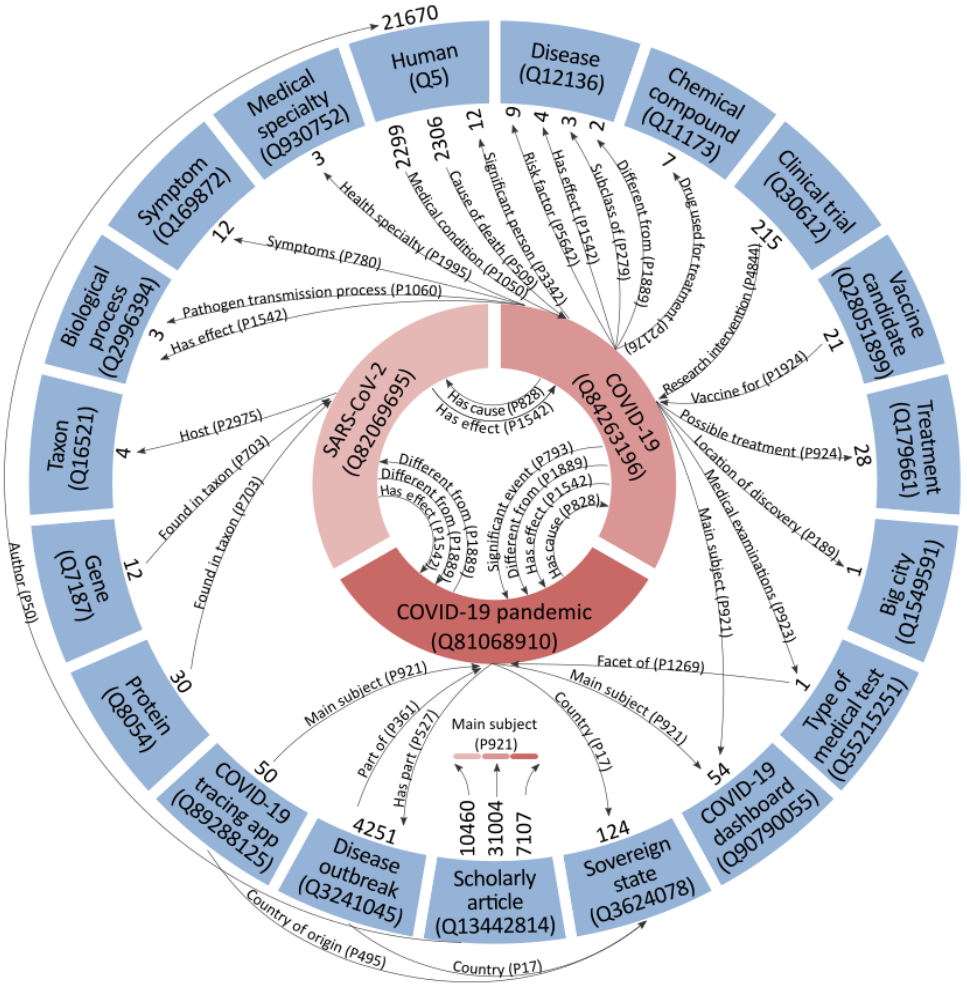

Fig. 2. Simplified skeleton of the data model of COVID-19 information on Wikidata. The three main COVID-related items (the 'C3 items')[9] are represented in red, selected classes of items related to these are shown in blue, with the relations between them represented as arrows. The number of statements 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[10], available at: https://w.wiki/auD, license: CC BY 4.0.

[9] COVID and C3 stand for any subset of {*COVID-19* [Q84263196], *SARS-CoV-2* [Q82069695], *COVID-19 pandemic* [Q81068910]}.
[10] Source queries: https://w.wiki/Ypc, https://w.wiki/Ypd, https://w.wiki/Ype, https://w.wiki/Ypg, https://w.wiki/Yph, and https://w.wiki/Ypi.

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-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 out 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). The collaborative work to populate and curate this data has been largely accomplished by WikiProject COVID-19[11], launched in March 2020 [15]. This WikiProject itself has a Wikidata item [Q87748614], and items are linked to it using the property *on focus list of Wikimedia project* [P5008].

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 [15, 23]. 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 [17, 18]. In Wikidata, several initiatives such as WikiCite for scholarly information [35-38] and Gene Wiki for genomic data [39] 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 [40]. Wikidata items are assigned their identifiers in external databases, including semantic resources, using human efforts and tools such as Mix'n'match [41]. 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 [39] and federation [41] in the context of the linked open data cloud [42]. 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. Examples of Wikidata properties used to define external identifiers can be found in Table 1.

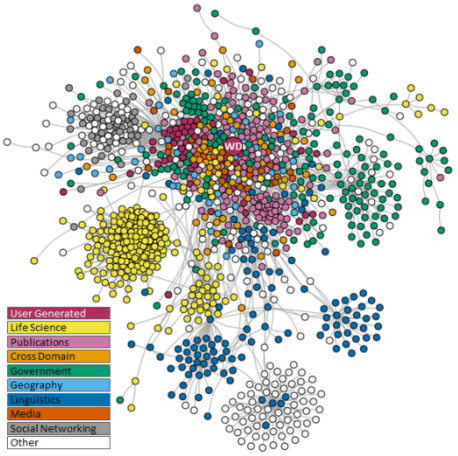

Fig. 3. Wikidata in the Linked Open Data Cloud. Databases indicated as circles (with Wikidata indicated as 'WD'), with grey lines linking databases in the network if their data is aligned, source dataset last updated May 2020 (available at: https://w.wiki/bYM, license: CC BY 4.0).

[11] https://www.wikidata.org/wiki/Wikidata:WikiProject_COVID-19

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] 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 $\underline{nt_Z} = t_Z - t_{Z-1}$
- 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 [23]. Later, the developed semantic database for the pandemic is checked by multiple layers of validation. Methods include RDF triples defining conditions for the usage of Wikidata properties, RDF validation schemas built in Shape Expressions (ShEx) to verify the structural accuracy of the statement of an item included in a given Wikidata class, and logical constraints implemented in SPARQL to verify the consistency of relational and non-relational claims in Wikidata as well as several tools based on edit history of Wikidata such as ORES to identify and eliminate database vandalism [43]. 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 [19], the infrastructure developed for the validation of RDF data in Wikidata helps assure a high level of consistency of the Wikidata knowledge graph.

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

[12] https://datahub.io/core/covid-19

[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

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

2.2. Multilingual representation

Thanks to its multilingual and language-independent data model as well as its link with various biomedical ontologies and knowledge bases, Wikidata's biomedical language coverage in English, French, German and Dutch is comparable to other semantic resources such as SNOMED-CT[16], BabelMeSH[17], and ICD-10[18] [23]. Despite the recent origin of the COVID-19 pandemic, Wikidata's coverage on the matter is already quite granular, with the main three COVID items linked to 17,000 other items via 55,000 relations at the time of writing. The degree of translation of that information is interestingly high with an important representation of the concepts in more than 50 languages (Fig. 4E). In fact, 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. However, this coverage varies between languages, with English as 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 linguistic coverage is less uneven than other biomedical semantic resources (e.g. SNOMED-CT and BabelMeSH) [45, 46] and is in line with efforts of generating multilingual language resources to be used for natural language processing purposes in clinical contexts [47].

The better coverage of English is explained in part by the higher support of this language in both biomedical language resources [48] and Wikipedia [49]. Cooperation with publishers such as Cochrane has a significant effect on English Wikipedia coverage, too [50]. The significant coverage of other languages like French, Spanish, German, Chinese

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

[17] 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.

[18] ICD-10: International Classification of Diseases, 10th Revision [44]: 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.

and Swedish in Medical Wikidata fits with their support by major biomedical multilingual databases: ICPC-2 [51] supports 24 languages[19], SNOMED-CT supports 7 languages, LOINC[20] supports 13 languages, BabelMeSH [52] supports 13 languages, and ICD-10 supports 42 languages.

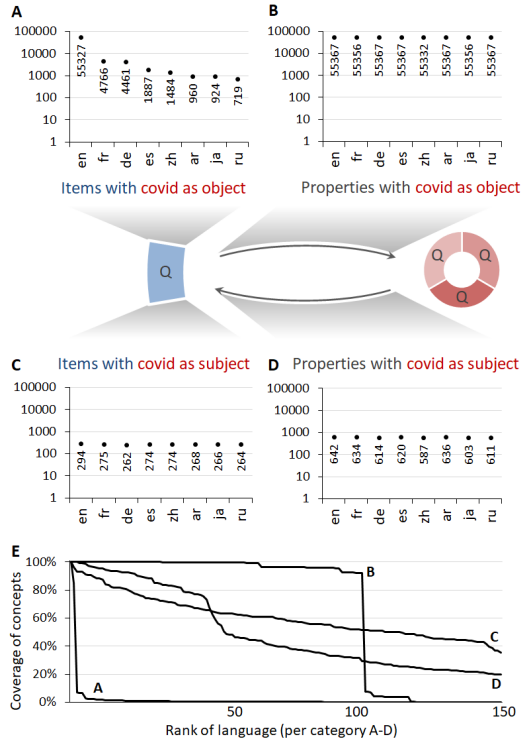

Fig. 4. Language representation of COVID-19-related statements. A-D) Language coverage for items and properties used in statements when either the object or subject is one of the three COVID-related items (as per Figure 2; note: log y-axis). The eight most common languages in Wikidata are shown: en=English, 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)

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[21] [53] and by the involvement of research institutions and scientists speaking these languages, particularly German and Dutch, in adding biomedical information to Wikidata [54, 55]. 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*[22] and others [28].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 [24]. However, there also appears to be an impact from the overall number of Wikidata labels for each language [25] and to the number of medical Wikipedia articles in each language [56] (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 10s 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 .65 to .97, median = .84, Supplementary Table S4), statistically significant even following stringent Bonferroni correction. To account for skew and data spanning multiple orders of magnitude, log10-transformed data was used for subsequent analyses. Pearson's correlation coefficients between all variables was high (Figure 5). A principal component analysis for the 90

[19] 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.

[20] 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.

[21] 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.

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

languages with complete data on all 7 indicators found that a single component explained 81% of the variance, with loadings ranging from .80 to .95. The smallest PCA loading and Spearman's correlation was for the number of viewers, which though still a strong association, was less correlated than the other variables by a substantial margin.

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.

	Medical Wikipedia, 2013 [56]		Wikidata labels, 2017 [25]		Population, 2019 [57]		Wikidata users, 2018 [24]
Rank	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

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.

	Medical Wikipedia articles https://w.wiki/Z6a		Medical Wikidata labels https://w.wiki/Z6h		Wikipedia and Wikidata users https://w.wiki/Z6W	
Rank	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

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 confirmed 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 [57] and the maximum human development index for countries where that language is an official language [58]. 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 similar positive skew to earlier data, so was also log10-transformed. Even though these analyses are necessarily exploratory, maximum development correlated more strongly than did the number of speakers (Figure 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 [59].

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 [25]. Consequently, encouraging the contribution by speakers of under-resourced and unrepresented languages to medical Wikipedia projects[23] and to Medical Wikidata is highly valuable to ameliorate and increase the language coverage of Wikidata as well as culturally appropriate contextualizations in medical and other domains.

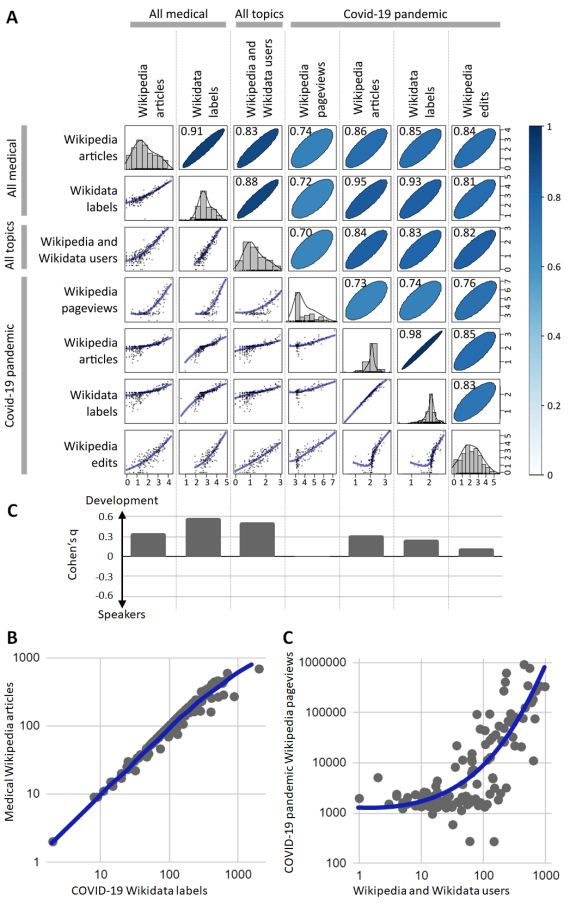

Fig. 5. A) All-versus-all pairwise correlations of log10-transformed values of seven metrics for 307 languages (data from sources detailed in tables 3 and 4). Histograms on diagonal indicate skew, scatter plots below diagonal indicate data and trendlines, ellipsoids above diagonal indicate Spearman's 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].

[23] 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 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.

	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	
Rank	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. Database alignment

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 [43].

The alignment of Wikidata entities to other entries on different databases is a collaborative process which, as 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 for locating the database ID. 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 are surfacing regularly, e.g. invalid DOIs stated in PubMed and PMC Europe[24]. 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[25] out of 7877[26] 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 [20, 17, 60]. 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 [20, 61] and for the integration of scholarly knowledge [62].

In the circumstances of the COVID-19 outbreak, a SPARQL query[27] 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 genes, of proteins, and of other items related to the ongoing COVID-19 pandemic in Wikidata. This confirms the centrality of

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

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

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

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

Wikidata within the linked open data cloud (cf. Fig. 3 and [42]) 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.

As shown in Table 5, 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. Most of these identifiers are added thanks to WikiProject WikiCite aiming to add support of bibliographic information on Wikidata [35-37]. 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[28] 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[29], 110323 COVID-19 full texts accessible on PubMed Central[30], 296450 COVID-19 publications on Dimensions[31], 211000 records on Semantic Scholar[32], 4778 records at ClinicalTrials.gov[33], 3295 records on arXiv ID[34], and 183 records on NIOSHTIC-2[35] 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 [37, 38]. 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.

Table 5

Main Wikidata properties used to represent the external identifiers of scholarly articles and clinical trials (as of August 31, 2020).

Wikidata ID	Wikidata Property	Count
P356	DOI	45101
P698	PubMed ID	42294
P6179	Dimensions Publication ID	16944
P932	PMCID	12590
P8150	COVIDWHO ID	11718
P8299	Semantic Scholar corpus ID	4612
P3098	ClinicalTrials.gov Identifier	246
P818	arXiv ID	47
P2880	NIOSHTIC-2 ID	23

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[36], and UMLS[37] as well as in open lexical databases like OBO Foundry ontologies (e.g. Human Phenotype Ontology) and Freebase as shown in Table 6. 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 [63]. 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 [64]. Items about diseases and symptoms are also aligned to several online encyclopedias (e.g. eMedicine, Encyclopedia Britannica, and

[28] https://scholia.toolforge.org/
[29] https://pubmed.ncbi.nlm.nih.gov/?term=COVID-19
[30] https://www.ncbi.nlm.nih.gov/pmc/?term=COVID-19
[31] https://tinyurl.com/y6kwrdth
[32] https://www.semanticscholar.org/search?q=COVID-19&sort=relevance
[33] https://clinicaltrials.gov/ct2/results?cond=COVID-19&term=&cntry=&state=&city=&dist=
[34] https://arxiv.org/search/?query=COVID-19&searchtype=all&source=header
[35] https://www2a.cdc.gov/nioshtic-2/Buildqyr.asp?S1=COVID-19&Submit=Search

[36] International Classification of Diseases, Tenth Revision (https://www.who.int/classifications/icd/en/)
[37] Unified Medical Language System (https://www.nlm.nih.gov/research/umls/index.html)

MedlinePlus) and to non-medical databases such as scholarly repositories (e.g. JSTOR[38]) and bibliographic databases (e.g. Microsoft Academic[39]) using external identifiers' statements. This can be explained by the efforts of WikiProject Source Metadata[40] 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 [35]. 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 [64] and is a rather established practice to provide Wikidata users with pointers to further specialized information pertaining to a given Wikidata item [65] and to allow comparison of medical data quality between Wikipedia and other encyclopedias [56].

Table 6

Main Wikidata properties used to represent the external identifiers of diseases and symptoms (as of August 31, 2020).

Wikidata ID	Wikidata Property	Diseases count	Symptoms count
P672	MeSH tree code	40	12
P2892	UMLS CUI	38	11
P494	ICD-10	32	8
P4229	ICD-10-CM[41]	32	1
P3827	JSTOR topic ID	32	10
P6366	Microsoft Academic ID	29	11
P493	ICD-9[42]	26	5
P673	eMedicine ID	24	2
P1417	Encyclopedia Britannica Online ID	23	7
P486	MeSH descriptor ID	23	9
P646	Freebase ID	21	10
P3841	Human Phenotype Ontology ID	18	9
P604	MedlinePlus ID	19	9
P508	BNCF[43] Thesaurus ID	17	7
P1296	Gran Enciclopedia Catalana ID	10	7
P8408	KBpedia[44] ID	16	7

The matching between Wikidata items and online encyclopedias and non-medical resources is not restricted to disease and symptoms. It additionally covers humans and sovereign states (Table 7) as well as films, computer applications and disease outbreaks (Table 8). The alignment to various metadata databases like VIAF[45], WorldCat[46], Library of Congress and IMDb[47] 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 [65, 66]. 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 [67]. Information about items in social media can also be retrieved to support the topic modelling of the coverage of the pandemic in social networks [68]. 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 [62].

[38] https://www.jstor.org/
[39] https://academic.microsoft.com/
[40] https://www.wikidata.org/wiki/WD:WikiProject_Source
[41] International Classification of Diseases, Tenth Revision, Clinical Modification
[42] International Classification of Diseases, Ninth Revision
[43] Biblioteca Nazionale Centrale di Firenze (Central National Library of Florence, Italy)
[44] https://kbpedia.org/
[45] Virtual International Authority File (http://viaf.org/)
[46] https://www.worldcat.org/
[47] Internet Movie Database (https://www.imdb.com/)

Table 7

Main Wikidata properties used to represent the external identifiers of humans and sovereign states (as of August 31, 2020).

Wikidata ID	Wikidata Property	Sovereign states	Humans
P214	VIAF ID	159	654
P7859	WorldCat Identities ID	146	548
P244	Library of Congress authority ID	125	458
P213	ISNI[48]	100	443
P646	Freebase ID	124	379
P2002	Twitter username	16	353
P227	GND[49] ID	125	308
P345	IMDb ID		274
P268	Bibliothèque nationale de France ID	177	269
P269	IdRef[50] ID	84	265
P998	DMOZ[51] ID	158	
P3417	Quora topic ID	141	73
P1417	Encyclopedia Britannica Online ID	138	53
P5400	GeoNLP ID	128	
P349	National Diet Library ID	127	54
P4801	LoC MARC[52] vocabularies ID	126	

Concerning drugs, proteins, genes and taxons, Wikidata items are mainly assigned external identifiers in the major knowledge graphs for pharmacology (e.g. MassBank[53]), for biodiversity (e.g. IRMNG[54]), for genomics (e.g. Entrez Gene) and for proteomics (e.g. PDB[55]) and are rarely linked to non-medical databases or to encyclopedias, as shown in Table 8. 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 [69]. The most important interest in matching these concepts in Wikidata to graph databases (e.g. Massbank, PDB, and KEGG[56]) and semi-structured databases (e.g. Guide to Pharmacology[57]) for bioinformatics rather than online encyclopedias is due to the better availability of genomic and proteomic information in these specialized semantic resources [64, 70]. The alignment of taxon items in Wikidata to biodiversity knowledge graphs (e.g. NCBI[58] 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 [71]. 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 [55] and Gene Wiki [39] and is enhanced in the context of the current pandemic through the contributions of WikiProject COVID-19 [15].

Table 8

Main Wikidata properties used to represent the external identifiers for other Wikidata classes (as of August 31, 2020).

Wikidata Class	Wikidata ID	Wikidata Property	Count
drug [Q11173]	P6689	MassBank accession ID	44
drug [Q11173]	P4964	SPLASH[59]	31
protein [Q8054]	P638	PDB structure ID	31
film [Q11424]	P345	IMDb ID	25

[48] https://isni.org/

[49] Gemeinsame Normdatei (German National Library, Germany), https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html

[50] Identifiants et Référentiels pour l'enseignement supérieur et la recherche (Identifiers and credentials for higher education and research in France)

[51] Directory Mozilla (https://dmoz-odp.org/)

[52] https://www.loc.gov/marc/

[53] https://massbank.eu/MassBank/

[54] Interim Register of Marine and Nonmarine Genera (https://www.irmng.org/)

[55] Protein Data Bank (https://www.rcsb.org/)

[56] Kyoto Encyclopedia of Genes and Genomes (https://www.genome.jp/kegg/)

[57] IUPHAR/BPS Guide to Pharmacology (https://www.guidetopharmacology.org/)

[58] National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/)

[59] Spectral Hash Identifier (https://splash.fiehnlab.ucdavis.edu/)

film [Q11424]	P2603	Kinopoisk film ID	23
film [Q11424]	P7177	Cinestaan film ID	22
disease outbreak [Q3241045]	P3984	subreddit	22
protein [Q8054]	P637	RefSeq[60] protein ID	18
committee group motion [Q97695005]	P8433	Swedish Riksdag document ID	18
film [Q11424]	P2529	ČSFD[61] film ID	17
drug [Q11173]	P267	ATC[62] code	17
protein [Q8054]	P352	UniProt protein ID	16
protein [Q8054]	P5458	Guide to Pharmacology Target ID	15
COVID-19 app [Q89288125]	P7771	PersonalData.IO ID	14
gene [Q7187]	P351	Entrez Gene ID	12
COVID-19 app [Q89288125]	P3418	Google Play Store app ID	12
gene [Q7187]	P2393	NCBI locus tag	11
macromolecular complex [Q22325163]	P7718	Complex Portal accession ID	11
protein fragment [Q78782478]	P638	PDB structure ID	11
drug [Q11173]	P231	CAS Registry[63] Number	9
drug [Q11173]	P715	DrugBank ID	9
drug [Q11173]	P665	KEGG ID	9
drug [Q11173]	P638	PDB structure ID	9
drug [Q11173]	P652	UNII[64]	9
protein [Q8054]	P705	Ensembl protein ID	8
COVID-19 app [Q89288125]	P3861	App Store app ID (global)	8
drug [Q11173]	P595	Guide to Pharmacology Ligand ID	8
drug [Q11173]	P6366	Microsoft Academic ID	8
disease outbreak [Q3241045]	P3479	Omni topic ID	7
taxon [Q16521]	P5055	IRMNG ID	6
taxon [Q16521]	P685	NCBI taxonomy ID	6

[60] NCBI Reference Sequence Database (https://www.ncbi.nlm.nih.gov/refseq/)

[61] Česko-Slovenská filmová databáze (Czech-Slovak Film Database, https://www.csfd.cz/)

[62] Anatomical Therapeutic Chemical (ATC) Classification System (https://www.whocc.no/atc_ddd_index/)

[63] https://www.cas.org/support/documentation/chemical-substances

[64] Unique Ingredient Identifier (https://fdasis.nlm.nih.gov/srs/)

3. Visualizing facets of COVID-19 via SPARQL

The flexible data model of Wikidata enables it to be highly multidisciplinary, including information ranging from medical to geopolitical to social aspects of the pandemic. Given the breadth of Wikidata's COVID-19-related information (examples in Supplementary Figure S1), extracting specific subsets of that information using SPARQL[65] can illustrate different aspects of the COVID-19 disease, its causative virus, and the resulting pandemic (extended list, Supplementary Table S1). Sample SPARQL queries for data visualizations commonly included in Wikidata-based COVID-19 dashboards are available at Supplementary Table S2 to show the variety of visualizations that can be generated using the Wikidata Query Service 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

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

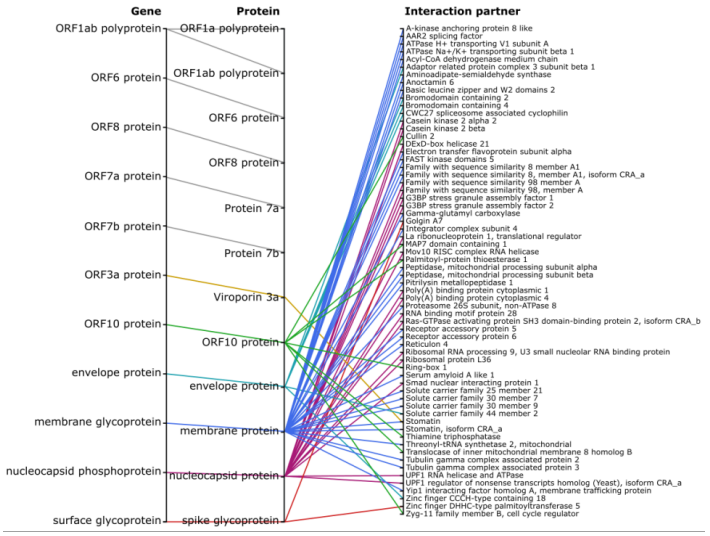

Fig. 6. SARS-CoV-2 interactions with the human proteome as of September 14, 2020 (available at: https://w.wiki/c3D, license: CC BY 4.0). Proteins encoded by SARS-CoV-2 genes (note that some genes encode multiple proteins) and the currently known human protein interaction partners (live data: https://w.wiki/beR).

which we have contributed substantially[66,67,68,69,70]. Several similar query collections exist, e.g. for COVID-19 in India[71].

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 [55, 73], 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) [64]. 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).

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

[66] 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

[67] SARS-CoV-2-Queries: extracts from the book "Wikidata Queries around the SARS-CoV-2 virus and pandemic" [72]; https://egonw.github.io/SARS-CoV-2-Queries/

[68] 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).

[69] Scholia queries: queries underlying COVID-19-related visualizations from the Wikidata-based scholarly profiling tool Scholia [35]; https://scholia.toolforge.org/

[70] 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

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

individuals deceased from COVID-19 (Figure 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.

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)

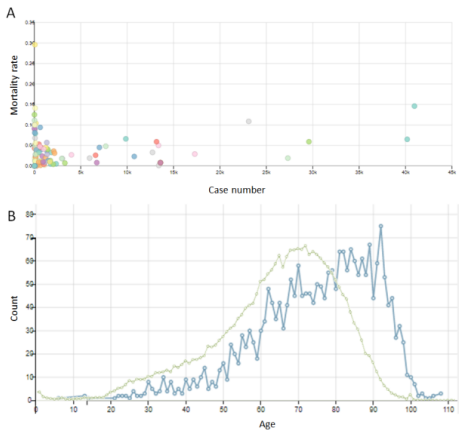

Fig. 8. Summary epidemiological data on the COVID-19 pandemic as of September 10, 2020 (available at: https://w.wiki/byW, license: CC BY 4.0). A) Correlation between the current number of cases and mortality rates in every country, calculated from numeric summary data for each region. Countries coloured 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).

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 [74] and clearly seen in other data sources, such that mortality peaks at the beginning of a disease outbreak [75]. 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 [76, 77].

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 [78]. Such statements allow broader querying of the literature as a network via co-occurrence of topics as the main subject of articles (Fig. 9B)[72]. This enables rapid traversal and faceting of the literature on topics in addition to the traditional links made by tracing citations [79], 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) [80].

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 [80, 81]. Further, Wikidata's rich biographical and institutional data makes extracting information on authors, institutions or others straightforward (Fig. S8), and eventually for other contributors too [82].

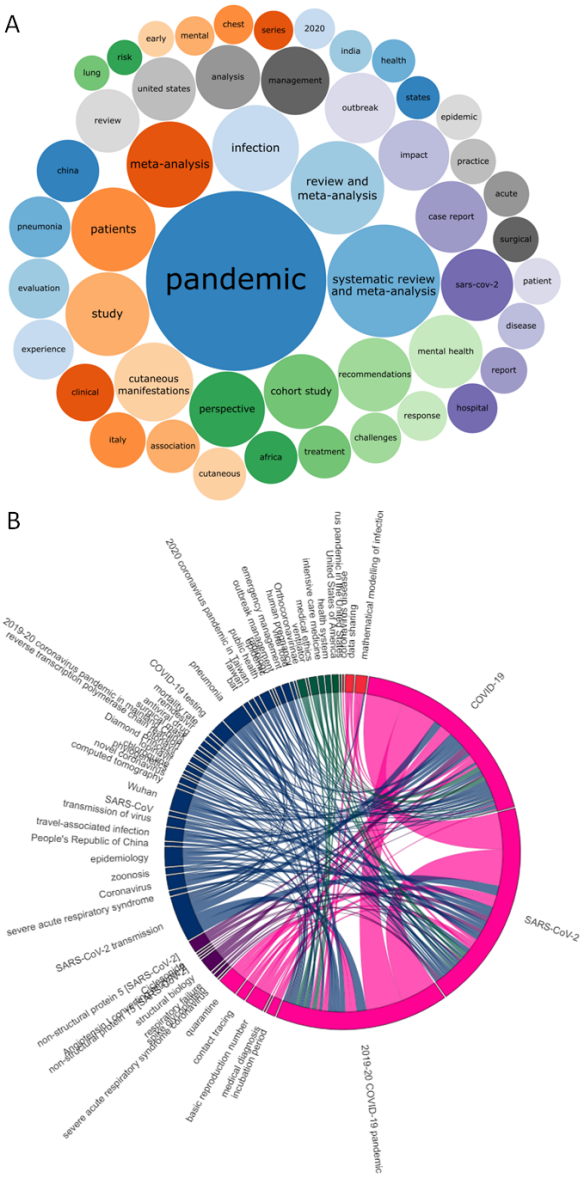

Fig. 9. COVID-19 publication topics as of September 10, 2020 (available at: https://w.wiki/byV, license: CC BY 4.0). A) Common words and word combinations (ngrams) in the titles of publications (live data: https://w.wiki/cFu). B) Co-occurrence of topics in publications with one of the COVID-related items 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).

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

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).

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		

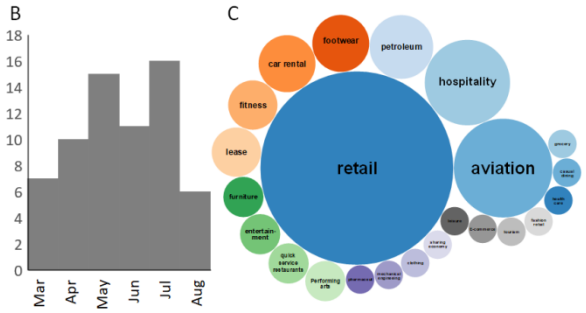

Fig. 10. Bankruptcies of publicly listed businesses due to the COVID-19 pandemic as of September 13, 2020 (available at: https://w.wiki/byY, license: CC BY 4.0). A) Tabular output of SPARQL query B) Bankruptcies per month C) ratios of different industries associated with bankrupt companies. (live data: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

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 [43].

4. Discussion

Many knowledge graphs have been recently developed to represent various types of COVID-19-related information, including government responses [5], epidemiology [8], clinical data [4], scholarly outputs and outcomes [9], economic impacts [10], physiopathology [2], social networking [11] among other features related to the COVID-19 pandemic. These semantic databases are mainly built using a combination of human efforts and crowdsourcing techniques [5]. 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 [7].

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) [3] and COVID-19 data hub (combining epidemiological data with social interactions) [12].

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 [14]. 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 [17].

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 amongst even other user-generated collaborative projects. Indeed, it becomes uniquely suited to highly dynamic topics such as the COVID-19 pandemic [15, 64]. 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 [83]. Wikidata is mainly enriched and adjusted by a community of over 25000 active human users on a daily basis[73] 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 [43]. By being highly multilingual, its human-readability extends well beyond English to support international contributions and reuse [23, 43]. 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 [23, 43]. One of the features of Wikidata is also providing hundreds of exemplary SPARQL queries[74], 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 [41]. 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 a richly interconnected and interdisciplinary network of >90 million items [23, 43]. 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 [3], COVID-19 data hub [12], COVID-19 Living Data[75] [84] and Knowledge4COVID-19[76] [85]) 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 [3] 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 [86]. 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 [3] is the possibility of integration of 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 to tasks related to COVID-19. However, application of CC0 on these databases causes them not to integrate information for semantic resources and datasets with 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 [87, 88]. This situation is similar to the status of regular group O red blood cells as a universal donor but restricted recipient [89].

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[77] and naval[78] 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

[73] https://www.wikidata.org/wiki/Special:Statistics
[74] https://w.wiki/pGw
[75] https://covid-nma.com/
[76] https://devpost.com/software/covid-19-kg

[77] https://en.wikipedia.org/wiki/COVID-19_pandemic_on_cruise_ships
[78] https://en.wikipedia.org/wiki/COVID-19_pandemic_on_naval_ships

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 [90, 91], systematically exposing the content of Wikidata to many eyes through its reuse in Wikipedia and SPARQL-based tools such as Scholia and COVID dashboards [15, 35], and using knowledge graph learning techniques to update items directly from textual databases like scholarly publications [92] and electronic health records [93]. Moreover, collaborative editing can cause several inaccuracies in the declaration of statements in open knowledge graphs disregarding the metadata standards of the knowledge bases [94]. 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[79] and can consequently cause matters about the trustworthiness of the reuse of data [94]. 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 [17]. Greater consistency of structure and accuracy is partly due to the involvement of more contributors in Wikidata than in other open knowledge graphs [17]. But it also stems from importing data from other rapidly-updated and curated databases (mainly from the linked open data cloud [23]) and from verification by overlapping methods (e.g. ShEx schemas[80], SPARQL-based logical constraints and bot edits [43, 95]). The data validation infrastructure of Wikidata seems to be in accordance with the latest updates in knowledge graph evaluation and refinement techniques [96, 97] and explains in part the reasons behind the robustness of the data model of COVID-19 information in this open knowledge graph.

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

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

5. Conclusion

In this research paper, we demonstrate the ability of open and collaborative knowledge graphs such as Wikidata to represent and integrate a large number of the multidisciplinary aspects of the COVID-19 information and to use SPARQL to generate summary visualizations about the infectious disease, the underlying pathogen, the resulting pandemic and related topics. We have shown how the community-driven and not centrally coordinated approach to editing 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. As an open semantic resource in the RDF format, Wikidata has become a hub for COVID-19 knowledge. 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 [64].

An important aspect of Wikidata's FAIRness is the Wikidata SPARQL query service (https://query.wikidata.org) [64]. 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 [35]. As shown here, SPARQL visualizations are an entrypoint for deeper insights into COVID-19, both regarding the biomedical facets of this still new disease, as well as into the societal details of the pandemic.

Another partner for FAIRness is user-friendly programmatic data access. Wikidata database dumps are available for download and local processing (https://www.wikidata.org/wiki/Wikidata:Database_download) in RDF, JSON and XML formats. 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/), arguably exposing the content even further.

Even though Wikidata is rich in COVID-19 knowledge, there is always room for future 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.

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.

6. Author statements

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 [98]. 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[81] 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 [99].

[81] 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.

7. 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] 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), pp. 779-788. doi:10.1016/j.dsx.2020.05.035.

[2] 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.

[3] 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), p. 181. doi:10.1038/s41597-020-0523-6.

[4] 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)., p. 136. doi:10.1038/s41597-020-0477-8.

[5] 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), p. 285. doi:10.1038/s41597-020-00609-9.

[6] 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), p. 2020030413. doi:10.20944/preprints202003.0413.v1.

[7] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, et al. CORD-19: The Covid-19 Open Research Dataset.

arXiv preprint arXiv:2004.10706 (2020). https://www.ncbi.nlm.nih.gov/pmc/articles/pmc7251955/.

[8] B. Xu, M. U. Kraemer, and Data Curation Group. Open access epidemiological data from the COVID-19 outbreak. *The Lancet Infectious Diseases*, 20.5 (2020), pp. 534. doi:10.1016/S1473-3099(20)30119-5.

[9] 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.

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

[11] 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. *arXiv preprint arXiv:2006.14492* (2020).

[12] E. Guidotti and D. Ardia. COVID-19 data hub. *Journal of Open Source Software*, 5.51 (2020), p. 2376. doi:10.21105/joss.02376.

[13] 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.

[14] D. Kagan, J. Moran-Gilad, and M. Fire. Scientometric trends for coronaviruses and other emerging viral infections. *GigaScience*, 9.8 (2020), p. giaa085. doi:10.1093/gigascience/giaa085.

[15] 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), p. 12. doi:10.1101/2020.04.05.026336.

[16] RDA COVID-19 Working Group. *Recommendations and Guidelines on data sharing*. Research Data Alliance, 2020. doi:10.15497/rda00052.

[17] M. Färber, F. Bartscherer, C. Menne, and A. Rettinger. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semantic Web, 9.1 (2018), pp. 77–129. doi:10.3233/SW-170275.

[18] 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.

[19] 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.

[20] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57.10 (2014), pp. 78-85. doi:10.1145/2629489.

[21] D. Jemielniak. *Common knowledge?: An ethnography of Wikipedia*, Stanford: Stanford University Press, 2014. ISBN:978-0804789448.

[22] P. Konieczny. Adhocratic governance in the Internet age: A case of Wikipedia. *Journal of Information Technology & Politics*, 7.4 (2010), pp. 263-283. doi:10.1080/19331681.2010.489408.

[23] H. Turki, T. Shafee, M. A. Hadj Taieb, M. Ben Aouicha, D. Vrandečić, D. Das, and H. Hamdi. Wikidata: A large-scale collaborative ontological medical database. *Journal of Biomedical Informatics*, 99 (2019), p. 103292. doi:10.1016/j.jbi.2019.103292.

[24] 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.

[25] L. A. Kaffee, A. Piscopo, P. Vougiouklis, E. Simperl, L. Carr, and L. Pintscher. A glimpse into babel: An analysis of multilinguality in Wikidata, in: *Proceedings of the 13th International Symposium on Open Collaboration*, ACM, 2017, p. 14. doi:10.1145/3125433.3125465.

[26] 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.

[27] 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.

[28] J. Samuel. Towards understanding and improving multilingual collaborative ontology development in Wikidata, in: *Companion of the The Web Conference 2018 on The Web Conference*, ACM, 2018, pp. 23-27.

[29] 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), p. baaa015. doi:10.1093/database/baaa015.

[30] A. Kirk. *Data visualisation: A handbook for data driven design*, Sage, 2016. ISBN:9781473966314.

[31] P. Hitzler and K. Janowicz. Linked Data, Big Data, and the 4th Paradigm. *Semantic Web*, 4.3 (2013), pp. 233-235. doi:10.3233/SW-130117.

[32] 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), p. 30. doi:10.1007/s13278-018-0507-0.

[33] 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, p. Paper 1.

[34] 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), p. e0144578.

[35] 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.

[36] D. Mietchen. State of WikiCite in 2020, in: *Workshop on Open Citations and Open Scholarly Metadata 2020*, 2020. doi:10.5281/zenodo.4019954.

[37] 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.

[38] D. Taraborelli, L. Pintscher, D. Mietchen, and S. Rodlund. WikiCite 2017 report. *Figshare* (2017). doi:10.6084/m9.figshare.5648233.

[39] 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), p. baw015. doi:10.1093/database/baw015.

[40] L. Ehrlinger and W. Wöß. Towards a Definition of Knowledge Graphs. *CEUR Workshop Proceedings*, 1695 (2016), pp. 1-4.

[41] 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.

[42] J. Debattista, C. Lange, S. Auer, and D. Cortis. Evaluating the quality of the LOD cloud: An empirical investigation. *Semantic Web*, 9.6 (2018), pp. 859-901. doi:10.3233/SW-180306.

[43] 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.

[44] 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), pp. 1105-1110. doi:10.1097/MLR.0b013e3181ef9d3e.

[45] 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.

[46] 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.

[47] 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.

[48] F. Freitas, S. Schulz, and E. Moraes. Survey of current terminologies and ontologies in biology and medicine. *Reciis*, 3.1 (2009), pp. 7-18. doi:10.3395/reciis.v3i1.239en.

[49] 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), pp. 1122-1129. doi:10.1136/jech-2016-208601.

[50] 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.

[51] 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.

[52] P. Fontelo, F. Liu, S. Leon, A. Abrahamane, and M. Ackerman. PICO Linguist and 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.

[53] 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), pp. 385-418. doi:10.1007/s10579-019-09453-9.

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

[55] 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), p. bax025. doi:10.1093/database/bax025.

[56] 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), p. e62. doi:10.2196/jmir.4069.

[57] D. M. Eberhand, G. F. Simons, and C. D. Fennig. *Ethnologue: Languages of the World*, SIL International, Dallas, Texas, 2020.

[58] 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.

[59] J. Cohen. *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum, 1988. ISBN:978-0-8058-0283-2.

[60] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, and D. Vrandečić (2014). 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.

[61] 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.

[62] 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), p. e7573. doi:10.3897/rio.1.e7573.

[63] A. Rahimi, T. Baldwin, and K. Verspoor. WikiUMLS: Aligning UMLS to Wikipedia via Cross-lingual Neural Ranking. *arXiv preprint arXiv:2005.01281* (2020).

[64] 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), p. e52614. doi:10.7554/eLife.52614.

[65] M. Klein and A. Kyrios. VIAFbot and the integration of library data on Wikipedia. *Code4lib journal*, 22 (2013).

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

[67] 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.

[68] 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), pp. 322-330. doi:10.1016/j.jbusres.2020.06.012.

[69] C. Ordun, S. Purushotham, and E. Raff. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082* (2020).

[70] 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), p. e175.

[71] R. Page. Towards a biodiversity knowledge graph. *Research Ideas and Outcomes*, 2 (2016), p. e8767. doi:10.3897/rio.2.e8767.

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

[73] M. Manske, U. Böhme, C. Püthe, and M. Berriman. GeneDB and Wikidata. *Wellcome open research*, 4 (2019), p. 114. doi:10.12688/wellcomeopenres.15355.2.

[74] 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), pp. 133-138. doi:10.1007/s13167-020-00207-0.

[75] 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), p. 147. doi:10.1016/j.jinf.2020.03.018.

[76] 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.

[77] B. C. Keegan and C. Tan. A Quantitative Portrait of Wikipedia's High-Tempo Collaborations during the 2020 Coronavirus Pandemic. *arXiv preprint arXiv:2006.08899* (2020).

[78] 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), pp. 217-218. doi:10.1016/j.jbi.2018.05.011.

[79] X. Hu, R. Rousseau, and J. Chen. On the definition of forward and backward citation generations. *Journal of Informetrics*, 5.1 (2011), pp. 27-36. doi:10.1016/j.joi.2010.07.004.

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

[81] 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), pp. e627-e630. doi:10.1016/S2214-109X(20)30113-3.

[82] 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.

[83] 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), pp. 167-195. doi:10.3233/SW-140134.

[84] 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), p. 108. doi:10.1186/s13643-020-01371-0.

[85] 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. *arXiv preprint arXiv:2008.07176* (2020).

[86] A. S. Dadzie, and M. Rowe. Approaches to visualising linked data: A survey. *Semantic Web*, 2.2 (2011), pp. 89-124. doi:10.3233/SW-2011-0037.

[87] 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), p. 127. doi:10.3897/zookeys.150.2189.

[88] 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), p. e12431. doi:10.3897/rio.3.e12431.

[89] D. M. Lublin. Universal RBCs. *Transfusion*, 40.11 (2000), pp. 1285-1289. doi:10.1046/j.1537-2995.2000.40111285.x.

[90] 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.

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

[92] 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), pp. 83-92. doi:10.1016/j.jbi.2018.03.011.

[93] 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), pp. 1-11. doi:10.1038/s41598-017-05778-z.

[94] 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), p. 188. doi:10.1038/s41597-020-0524-5.

[95] 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.

[96] 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), pp. 63-93. doi:10.3233/SW-150175.

[97] H. Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8.3 (2017), pp. 489-508. doi:10.3233/SW-160218.

[98] 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.

[99] 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.

Supplementary Data

Supplementary tables

Table S1. List of the tasks fulfilled by the SPARQL queries for the visualization of the COVID-19 information in Wikidata

Task	Description
Genomic data and clinical knowledge	
Z1	Symptoms of COVID-19 (SPEED, SARS-CoV-2-Queries)
Z2	Potential treatments of COVID-19 (SPEED)
Z3	Linnean Taxonomy of SARS-CoV-2 (SPEED)
Z4	All SARSr viruses (SARS-CoV-2-Queries)
Z5	Coronaviruses that infect humans (SARS-CoV-2-Queries)
Z6	All betacoronaviruses (SARS-CoV-2-Queries, WPCOVID)
Z7	All coronaviruses (SARS-CoV-2-Queries)
Z8	Comparing viruses with SARS-CoV-2 (SARS-CoV-2-Queries)
Z9	NCBI Taxonomy IDs of coronaviruses (SARS-CoV-2-Queries)
Z10	SARS-CoV-2 genomes (SARS-CoV-2-Queries)
Z11	SARS-CoV-2 genes (SARS-CoV-2-Queries)
Z12	SARS-CoV-2 proteins (SARS-CoV-2-Queries)
Z13	SARS-CoV-2 protein complexes (SARS-CoV-2-Queries)
Z14	SARSr genes (SARS-CoV-2-Queries)
Z15	SARSr proteins (SARS-CoV-2-Queries)
Z16	Human coronavirus' genes (SARS-CoV-2-Queries)
Z17	Human coronavirus' proteins (SARS-CoV-2-Queries)
Z18	Coronavirus' proteins interacting with human proteins (SARS-CoV-2-Queries)
Z19	Biological process for the pathogenesis of coronaviruses (SARS-CoV-2-Queries)
Z20	Antibodies for the coronaviruses (SARS-CoV-2-Queries)
Z21	Vaccines for the coronaviruses (SARS-CoV-2-Queries)
Z22	Drugs for the coronaviruses (SARS-CoV-2-Queries)
Z23	COVID-19, COVID-19 pandemic and SARS-CoV-2 in the context of the Wikidata knowledge graph (Scholia)
Epidemiology	
Z24	Daily evolution of the global number of COVID-19 cases (SARS-CoV-2-Queries, WPCOVID, COVID-19 Summary)
Z25	Daily evolution of the number of COVID-19 Cases by Country (SPEED)
Z26	Daily evolution of the number of COVID-19 Deaths by Country (SPEED)
Z27	Daily evolution of the COVID-19 Mortality Rate by Country (SPEED)
Z28	Daily evolution of the number of COVID-19 Clinical Tests by Country (SPEED)
Z29	Daily evolution of the COVID-19 Positive Test Rate by Country (SPEED)
Z30	Daily evolution of the number of COVID-19 Recoveries by Country (SPEED)
Z31	Daily evolution of the COVID-19 Recovery Rate by Country (SPEED)
Z32	Daily evolution of the number of COVID-19 Cases in a given country (SPEED, SARS-CoV-2-Queries)
Z33	Daily evolution of the number of COVID-19 Deaths in a given country (SPEED, SARS-CoV-2-Queries)
Z34	Daily evolution of the number of COVID-19 Clinical Tests in a given country (SPEED)
Z35	Daily evolution of the number of COVID-19 Recoveries in a given country (SPEED)
Z36	Daily evolution of the COVID-19 Mortality Rate in a given country (SPEED)
Z37	Daily evolution of the COVID-19 Positive Clinical Test Rate in a given country (SPEED)
Z38	Daily evolution of the COVID-19 Recovery Rate in a given country (SPEED)
Z39	Daily evolution of the number of COVID-19 Cases by administrative subdivision of a given country (SPEED)
Z40	Daily evolution of the number of COVID-19 Deaths by administrative subdivision of a given country (SPEED)
Z41	Daily evolution of the COVID-19 Mortality Rate by administrative subdivision of a given country (SPEED)

Z42	Daily evolution of the number of COVID-19 New Cases (SPEED)
Z43	Daily evolution of the number of COVID-19 New Deaths (SPEED)
Z44	Daily evolution of the number of COVID-19 New Clinical Tests (SPEED)
Z45	Daily evolution of the number of COVID-19 New Recoveries (SPEED)
Z46	Daily evolution of the number of COVID-19 Active Cases (SPEED)
Z47	Daily evolution of the number of COVID-19 Clinical Tests by Laboratory in a given country (SPEED)
Z48	Number of COVID-19 Cases by administrative subdivision of a given country (SPEED)
Z49	Number of COVID-19 Deaths by administrative subdivision of a given country (SPEED)
Z50	COVID-19 Mortality Rate by administrative subdivision of a given country (SPEED)
Z51	Number of COVID-19 Cases per Capita by administrative subdivision of a given country (SPEED)
Z52	Number of COVID-19 Deaths per Capita by administrative subdivision of a given country (SPEED)
Z53	Number of COVID-19 Cases per Area by administrative subdivision of a given country (SPEED)
Z54	Number of COVID-19 Deaths per Area by administrative subdivision of a given country (SPEED)
Z55	Current Epidemiological Status in a given country (SPEED)
Z56	Number of COVID-19 Clinical Tests by Laboratory in a given country (SPEED)
Z57	Map of Affected Countries (SPEED, WPCOVID)
Z58	Number of COVID-19 Cases by Country (SPEED, WPCOVID)
Z59	Number of COVID-19 Cases per 100000 inhabitants by Country (SPEED)
Z60	Number of COVID-19 Deaths by Country (SPEED)
Z61	Number of COVID-19 Deaths per 100000 inhabitants by Country (SPEED)
Z62	COVID-19 Mortality rates by Country (SPEED)
Z63	Number of COVID-19 Clinical Tests by Country (SPEED)
Z64	Number of COVID-19 Clinical Tests per 100000 inhabitants by Country (SPEED)
Z65	Number of COVID-19 Recoveries by Country (SPEED)
Z66	Number of COVID-19 Recoveries per 100000 inhabitants by Country (SPEED)
Z67	Famous COVID-19 Victims (SPEED, WPCOVID, COVID-19 Summary)
Z68	Age distribution of Famous COVID-19 Victims (COVID-19 Summary)
Z69	Field of work of Famous COVID-19 Victims (COVID-19 Summary)
Z70	Place of birth of Famous COVID-19 Victims (COVID-19 Summary)
Z71	Number of COVID-19 Cases per area by Country (SPEED, COVID-19 Summary)
Z72	Number of COVID-19 Deaths per area by Country (SPEED)
Z73	Number of COVID-19 Clinical Tests per area by Country (SPEED)
Z74	Number of COVID-19 Recoveries per area by Country (SPEED)
Z75	Number of COVID-19 Cases in function of the number of clinical tests in a given country (SPEED)
Z76	Number of COVID-19 Deaths in function of the number of cases in a given country (SPEED)
Z77	COVID-19 Mortality Rate in function of the number of cases in a given country (SPEED)
Z78	Number of COVID-19 cases in an administrative subdivision of a given country in function of population (SPEED)
Z79	Number of COVID-19 cases in an administrative subdivision of a given country in function of area (SPEED)
Z80	Number of COVID-19 cases in an administrative subdivision of a given country in function of population Density Rate (SPEED)
Z81	Number of COVID-19 deaths in an administrative subdivision of a given country in function of population (SPEED)
Z82	Number of COVID-19 deaths in an administrative subdivision of a given country in function of area (SPEED)
Z83	Number of COVID-19 deaths in an administrative subdivision of a given country in function of population Density Rate (SPEED)
Z84	COVID-19 Mortality Rate in an administrative subdivision of a given country in function of population (SPEED)
Z85	COVID-19 Mortality Rate in an administrative subdivision of a given country in function of area (SPEED)
Z86	COVID-19 Mortality Rate in an administrative subdivision of a given country in function of population Density Rate (SPEED)
Z87	Number of COVID-19 new cases in a given country in function of number of old cases (SPEED)
Z88	Global number of COVID-19 Cases in function of the global number of clinical tests (SPEED)
Z89	Global number of COVID-19 Deaths in function of the global number of cases (SPEED)
Z90	COVID-19 Global Mortality Rate in function of the global number of cases (SPEED)
Z91	Country-level number of COVID-19 Cases in function of Country Population (SPEED)
Z92	Country-level number of COVID-19 Cases in function of Country Area (SPEED)
Z93	Country-level number of COVID-19 Cases in function of Country Population Density Rate (SPEED)
Z94	Country-level number of COVID-19 Deaths in function of Country Population (SPEED)
Z95	Country-level number of COVID-19 Deaths in function of Country Area (SPEED)
Z96	Country-level number of COVID-19 Deaths in function of Country Density Rate (SPEED)
Z97	Country-level COVID-19 Mortality Rate in function of Country Population (SPEED)
Z98	Country-level COVID-19 Mortality Rate in function of Country Area (SPEED)

Z99	Country-level COVID-19 Mortality Rate in function of Country Population Density Rate (SPEED)
Z100	Duration between first case and first death based on number of cases and number of deaths in a given country (SARS-CoV-2-Queries)
Z101	Lockdowns due to the COVID-19 pandemic (WPCOVID)
Research outputs and computer applications	
Z102	Scholarly publications about COVID-19 pandemic and SARS-CoV-2 (SPEED, SARS-CoV-2-Queries, WPCOVID, Scholia)
Z103	Tools and Resources about COVID-19 pandemic by type (SPEED)
Z104	Tools and Resources about COVID-19 pandemic (SPEED)
Z105	Tools and Resources about COVID-19 pandemic by publisher (SPEED)
Z106	Tools and Resources about COVID-19 pandemic by license (SPEED)
Z107	Tools and Resources about COVID-19 pandemic by field of work (SPEED)
Z108	Clinical trials about COVID-19 pandemic (SARS-CoV-2-Queries)
Z109	Scholarly publications about the virus transmission of coronaviruses (SARS-CoV-2-Queries)
Z110	Scholarly publications about the SARS-CoV-2 genes (SARS-CoV-2-Queries)
Z111	Scholarly publications about the SARS-CoV-2 proteins (SARS-CoV-2-Queries)
Z112	Scholarly publications about coronaviruses (SARS-CoV-2-Queries)
Z113	Scholarly publications about human coronaviruses (SARS-CoV-2-Queries)
Z114	Contact tracing protocols related to the COVID-19 pandemic (WPCOVID)
Z115	Scholarly publications about COVID-19 pandemic and SARS-CoV-2 by year (Scholia)
Z116	Research scientists mostly publishing scholarly publications about COVID-19 pandemic and SARS-CoV-2 (Scholia)
Z117	Collaboration network of the research scientists working on COVID-19 pandemic and SARS-CoV-2 (Scholia)
Z118	Topics of the scholarly publications about COVID-19 pandemic and SARS-CoV-2 (Scholia)
Z119	Co-occurring topic graph of the scholarly publications about COVID-19 pandemic and SARS-CoV-2 (Scholia)
Z120	Map of cities and countries evocated by the scholarly publications about COVID-19 pandemic and SARS-CoV-2 (Scholia)
Z121	Research scientists mostly cited by the scholarly publications about COVID-19 pandemic and SARS-CoV-2 (Scholia)
Z122	Venues and series mostly publishing research works about the COVID-19 pandemic and SARS-CoV-2 (Scholia)
Z123	Most cited research publications about COVID-19 pandemic and SARS-CoV-2 (Scholia)
Z124	Map of institutions publishing research works about COVID-19 pandemic and SARS-CoV-2 (Scholia)
Z125	Citation network of research countries working on COVID-19 pandemic and SARS-CoV-2 (Scholia)
Z126	Awards received by authors who published on COVID-19 pandemic and SARS-CoV-2 (Scholia)
Z127	Scholarly publications about COVID-19 and SARS-CoV-2 with missing *main subject* [P921] values (SARS-CoV-2-Queries, WPCOVID)
Other	
Z128	Images from Wikimedia Commons about COVID-19 pandemic and SARS-CoV-2 (SPEED)
Z129	COVID-19 Factbook (SPEED)
Z130	Bankrupt businesses due to the COVID-19 pandemic (WPCOVID)
Z131	Properties used to model COVID-19 knowledge in Wikidata (WPCOVID)

Table S2. List of sample queries on COVID-19. The information contained therein is similar to visualizations in many stand-alone COVID-19 dashboards, covering an overview of COVID-19, international situation, international daily epidemiological evolution, Tunisian daily epidemiological evolution, Tunisian governorate-level situation, Tunisian correlations, and worldwide correlations. Each of the sheets has a Title column with a brief summary for each query and a URL column with a link to the live record on Wikidata.

Table available as Query/COVID-19.xlsx in http://doi.org/10.5281/zenodo.4022591.

Table S3. Raw data and correlation statistics for datasets summarised in tables 3 and 4, including Pearson's, Spearman's, and Cohen's coefficients for the raw data and Spearman's coefficients and principal component analysis of the log-10 transformed data.

Table available as docs/Fig5Corr/T3+4.xlsx in http://doi.org/10.5281/zenodo.4022591.

Table S4. Spearman's rho on raw data (pairwise) of untransformed variables from tables 3 and 4 against max development index for countries speaking each language as an official language, and number of native speakers. Final column indicates Cohen's q value (calculated as the difference between the Fisher-transformed Spearman's rho values i.e., $q = z'(r_{(\text{development,Wikidata med labels})}) - z'(r_{(\text{number of speakers, Wikidata med labels})})$), comparing these two for the stronger correlate for variables from tables 3 and 4. Positive values indicate max development index as the stronger correlate, while negative values would indicate number of native speakers as the stronger correlate. Differences of >.5 are considered "large" and unusual for the social sciences, .3 "medium" and .1 "small".

	Spearman's rho	Spearman's rho	**Cohen's q**
	Max development	Number speakers	**development - speakers**
Medical Wikipedia articles	.71	.48	**.36**
Medical Wikidata labels	.76	.38	**.59**
Wikipedia and Wikidata Users	.62	.21	**.51**
COVID19 pandemic Wikipedia pageviews	.53	.53	**.00**
COVID Wikipedia pages	.71	.52	**.31**
COVID Wikidata content	.69	.53	**.26**
COVID Wikipedia edits	.63	.55	**.12**

Supplementary figures

This section of the supplementary data includes additional array of visualizations that were not able to fit in the main text but that exemplify the diversity of additional valuable information that can be extracted out of the Wikidata knowledge base.

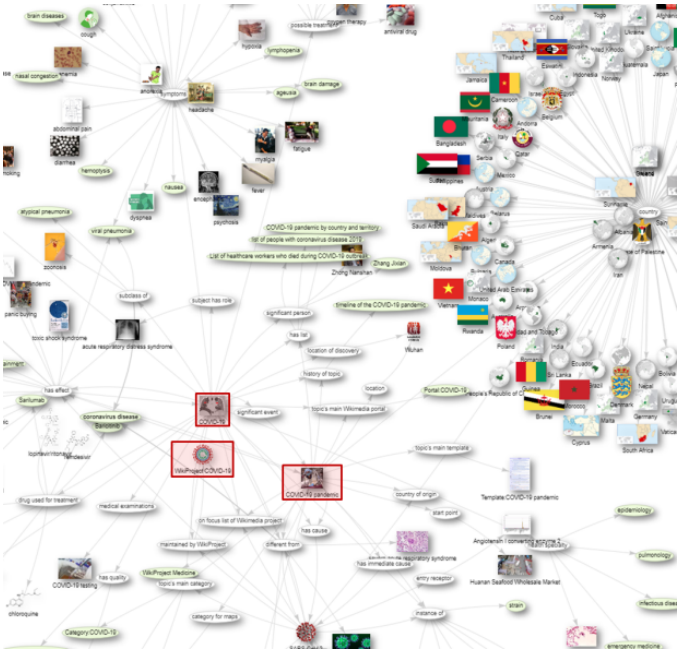

Fig. S1. Snapshot of the extended graph of the three main COVID items and the statements for which they are the subject. Linked items demonstrate the variety of topics for which the three main COVID items (indicated in red) are the subject and present a small subset of the classes indicated in Fig. 2. (Available at: https://w.wiki/cPa, live data: https://w.wiki/xYE, Access Date: August 19, 2020)

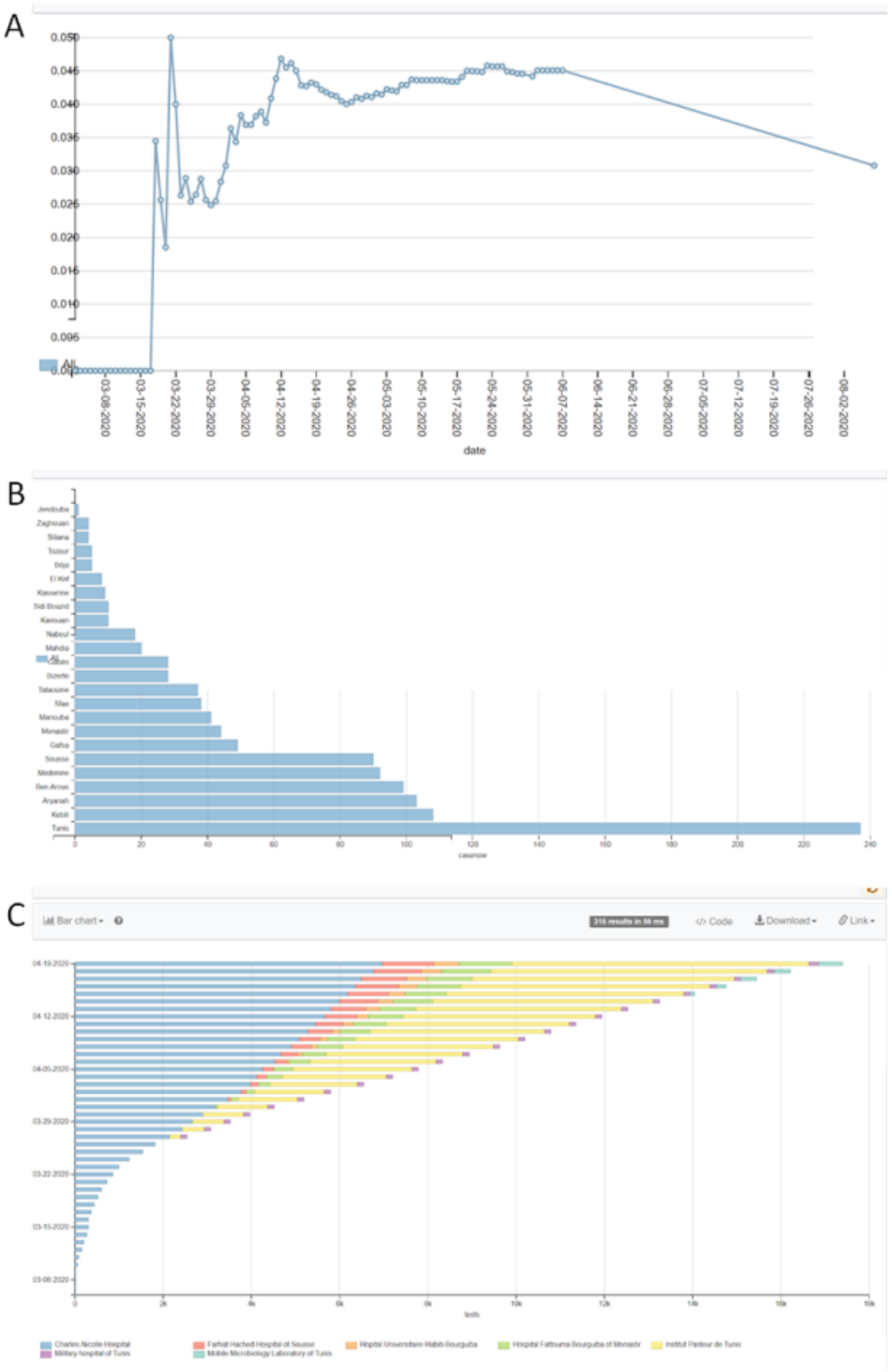

Fig. S2. Epidemiological data for Tunisia as of August 16, 2020. The SPEED website was set up as a COVID-19 data dashboard for Tunisia (Available at: https://w.wiki/cQC). A) Daily mortality rate from COVID-19 in Tunisia (live data: https://w.wiki/N2p). B) Tunisian governorate-level cases (live data: https://w.wiki/N9Y). C) Daily Evolution of Clinical tests by laboratory in Tunisia (live data: https://w.wiki/NEb).

A

dateOfDeath	name	citizenship	profession
16 August 2020	Chetan Chauhan	India	cricketer
14 August 2020	Moisés Mamani Colquehuanca	Peru	politician
13 August 2020	Darío Vivas	Venezuela	politician
13 August 2020	Gulnazar Keldi	Tajikistan	journalist
11 August 2020	Trini Lopez	United States of America	film actor
11 August 2020	Rahat Indori	India	lyricist
11 August 2020	Sixto Brillantes	Philippines	lawyer
9 August 2020	Kamala	United States of America	professional wrestler
9 August 2020	Tony Moussa	Syria	actor
8 August 2020	Alfredo Lim	Philippines	police officer
8 August 2020	Buruju Kashamu	Nigeria	politician

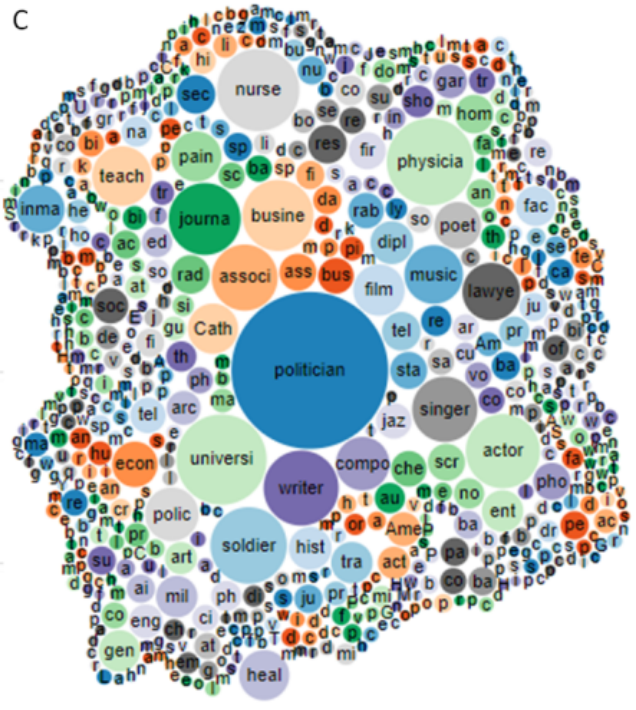

Figure S3. People listed in Wikidata deceased due to COVID-19 as of August 16, 2020 (Available at: https://w.wiki/cQK). A) As tabular output, ranked by date of death (live data: https://w.wiki/Mgv). B) Portrait images available under a CC BY-compatible license, ranked by how well-described the depicted individuals are in Wikidata (number of identifiers + statements + sitelinks) (live data: https://w.wiki/bzJ). C) as bubble diagram of professions (live data: https://w.wiki/bTz).

A

A pneumonia outbreak associated with a new coronavirus of probable bat origin
Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China
A Novel Coronavirus from Patients with Pneumonia in China, 2019
A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster
Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia
A new coronavirus associated with human respiratory disease in China
Structure of SARS coronavirus spike receptor-binding domain complexed with receptor.
Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation
Features and development of Coot
SARS and MERS: recent insights into emerging coronaviruses.

B

77	wd:Q63881333	Ralph S Baric	wd:Q91697408	Trypsin Treatment Unlocks Barrier for Zoonotic Bat Coronavirus Infection
73	wd:Q1079331	Christian Drosten	wd:Q83388131	The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health - The latest 2019 novel coronavirus outbreak in Wuhan, China
50	wd:Q77049459	Zihe Rao	wd:Q24817106	Design of wide-spectrum inhibitors targeting coronavirus main proteases
49	wd:Q55186759	Eric J Snijder	wd:Q36676674	Discovery of an essential nucleotidylating activity associated with a newly delineated conserved domain in the RNA polymerase-containing protein of all nidoviruses.
43	wd:Q37063445	Stefan Pöhlmann	wd:Q28730201	Influenza and SARS-coronavirus activating proteases TMPRSS2 and HAT are expressed at multiple sites in human respiratory and gastrointestinal tracts
41	wd:Q89552216	Michael Farzan	wd:Q28188496	Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus
39	wd:Q28152069	Edward C. Holmes	wd:Q34545653	Redefining the invertebrate RNA virosphere.
39	wd:Q84383569	Xiao-Shuang Zheng	wd:Q84367633	A pneumonia outbreak associated with a new coronavirus of probable bat origin
37	wd:Q88324943	Xu-Rui Shen	wd:Q84367633	A pneumonia outbreak associated with a new coronavirus of probable bat origin
37	wd:Q89108863	Barney S. Graham	wd:Q45323629	Safety, tolerability, and immunogenicity of two Zika virus DNA vaccine candidates in healthy adults: randomised, open-label, phase 1 clinical trials.

Figure S4. Partial citation network within Wikidata as of August 16, 2020 (Available at: https://w.wiki/cQV). The citation network around COVID-19 is currently rather incomplete but part of the larger, ongoing WikiCite project to represent all citation data within Wikidata as a fully open citation network. A) publications cited from C3 papers (live data: https://w.wiki/b$h) B) authors most frequently cited by C3 papers (live data: https://w.wiki/b$i).

count	venue	venueLabel	publisherLabel
2036	wd:Q58465838	medRxiv	Cold Spring Harbor Laboratory
1155	wd:Q546003	The BMJ	BMJ
823	wd:Q15716684	Journal of Medical Virology	Wiley-Blackwell
532	wd:Q19835482	bioRxiv	Cold Spring Harbor Laboratory
507	wd:Q5133764	Clinical Infectious Diseases	Oxford University Press
469	wd:Q939416	The Lancet	Elsevier
428	wd:Q6051382	International Journal of Environmental Research and Public Health	MDPI
420	wd:Q6295344	Journal of Infection	Elsevier
389	wd:Q1470970	Journal of the American Medical Association	American Medical Association
356	wd:Q15262334	International Journal of Infectious Diseases	Elsevier
347	wd:Q15766374	Dermatologic Therapy	Wiley-Blackwell
329	wd:Q582728	The New England Journal of Medicine	Massachusetts Medical Society
311	wd:Q6029185	Infection Control and Hospital Epidemiology	University of Chicago Press
274	wd:Q15724248	The Lancet Infectious Diseases	Elsevier

Figure S5. Most common publication venues for C3-themed papers (published and preprint) as of August 16, 2020. Even with Wikidata's currently incomplete coverage of articles hosted on preprint servers, they are clearly a significant location for COVID-related publications (Available at: https://w.wiki/cQX, live data: https://w.wiki/bd$).

Start date	Trial	Intervention	Sponsor
2020-05-12	Acalabrutinib Study With Best Supportive Care Versus Best Supportive Care in Subjects Hospitalized With COVID-19.		AstraZeneca
2020-05-10	COVID-19 Pneumonitis Low Dose Lung Radiotherapy (COLOR-19)		
2020-05-05	Levamisole and Isoprinosine in the Treatment of COVID19: A Proposed Therapeutic Trial	azithromycin	
2020-05-05	Levamisole and Isoprinosine in the Treatment of COVID19: A Proposed Therapeutic Trial	levamisole	
2020-05-05	Levamisole and Isoprinosine in the Treatment of COVID19: A Proposed Therapeutic Trial	hydroxychloroquine	
2020-05-05	Levamisole and Isoprinosine in the Treatment of COVID19: A Proposed Therapeutic Trial	inosine pranobex	
2020-04-24	Acalabrutinib Study With Best Supportive Care Versus Best Supportive Care in Subjects Hospitalized With COVID-19. CALAVI (Calquence Against the Virus)		AstraZeneca
2020-04-16	Austrian CoronaVirus Adaptive Clinical Trial (COVID-19)	candesartan	Medical University of Vienna
2020-04-16	Austrian CoronaVirus Adaptive Clinical Trial (COVID-19)	hydroxychloroquine	Medical University of Vienna
2020-04-16	Austrian CoronaVirus Adaptive Clinical Trial (COVID-19)	chloroquine	Medical University of Vienna

Fig. S6. Information regarding clinical trials on interventions to treat COVID-19 as of August 16, 2020 (Available at https://w.wiki/cQb, live data: https://w.wiki/bav)

toolLabel	COVID-19 European Dashboard
tool	wd:Q91219501
typeLabel	COVID-19 dashboard
URL	<https://qap.ecdc.europa.eu/public/extensions/COVID-19/COVID-19.html>
publisherLabel	European Centre for Disease Prevention and Control
licenseLabel	All Rights Reserved
toolLabel	COVID Racial Data Tracker
tool	wd:Q96655300
typeLabel	COVID-19 dashboard
toolLabel	COVID Atlas
tool	wd:Q96777164
typeLabel	COVID-19 dataset
toolLabel	COVID Atlas
tool	wd:Q96777164
typeLabel	COVID-19 search engine
toolLabel	Apturi Covid
tool	wd:Q97058482
typeLabel	COVID-19 app

Fig. S7. Computer applications and their types as of August 16, 2020 (Available at: https://w.wiki/cQg, live data: https://w.wiki/NVp)

A

count	award	awardLabel	recipients
4	wd:Q15631401	Fellow of the Royal Society	Bryan Grenfell, Malik Peiris, Edward C. Holmes, Gagandeep Kang
4	wd:Q24081923	Fellow of the Academy of Medical Sciences	Simon Wessely, Maria Zambon, Neil M. Ferguson, Clive Ballard
3	wd:Q7241433	Presidential Early Career Award for Scientists and Engineers	Russ Altman, John Brownstein, Namandjé N. Bumpus
3	wd:Q10762848	Officer of the Order of the British Empire	Bryan Grenfell, W. John Edmunds, Neil M. Ferguson
3	wd:Q26204035	Fellow of the Royal College of Physicians	Simon Wessely, Francine Ntoumi, Philip I. Murray
3	wd:Q59767813	Fellow of the American Institute for Medical and Biological Engineering	Russ Altman, Cato T. Laurencin, Elizabeth Krupinski
3	wd:Q63208574	Fellow of the African Academy of Sciences	Alimuddin Zumla, Abba Gumel, Francine Ntoumi
2	wd:Q5442484	AAAS Fellow	Ira Longini, Betz Halloran
2	wd:Q23697744	Kurt Lewin Medal	Alexander Haslam, Jolanda Jetten
2	wd:Q59771498	Fellow of the Academy of the Social Sciences in Australia	Helen Christensen, Jolanda Jetten
2	wd:Q59771619	Fellow of the Australian Academy of Health and Medical Sciences	Helen Christensen, Katherine Kedzierska
2	wd:Q61744587	Fellow of the American Statistical Association	Ira Longini, Betz Halloran
2	wd:Q72859645	Associate Fellow of the African Academy of Sciences	Cato T. Laurencin, George F. Gao

B

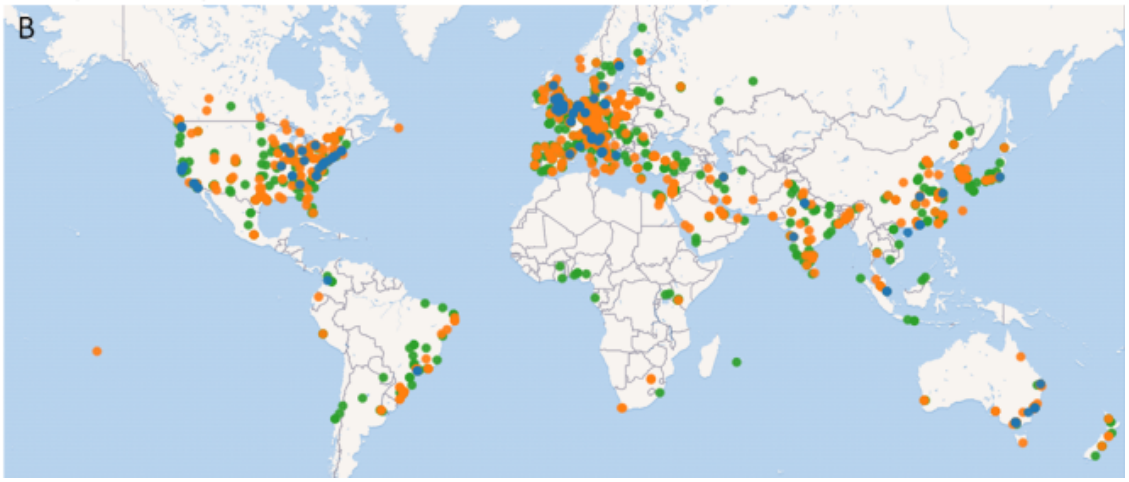

Fig. S8. Information on authors of articles on COVID-related topics as of August 16, 2020 (Available at: https://w.wiki/cQh). A) Awards most frequently received by authors of C3 papers (live data: https://w.wiki/ban), B) Map of organizations associated with works about C3 with institutions that have published a single paper on the topic in green, those that have published 1-10 in orange, and those having published >10 in blue (live data: https://w.wiki/cG4).

A

outbreak	label	URL
wd:Q89713663	2020 COVID-19 pandemic in the state of São Paulo	<https://www.seade.gov.br/coronavirus/>
wd:Q87743858	COVID-19 pandemic in Scotland	<https://www.gov.scot/coronavirus-covid-19/>
wd:Q87743873	2020 COVID-19 pandemic in Ohio	<https://coronavirus.ohio.gov/>
wd:Q87901408	2020 COVID-19 pandemic in Alberta	<https://www.alberta.ca/coronavirus-info-for-albertans.aspx>
wd:Q88097247	2020 COVID-19 pandemic in Gujarat	<https://gujcovid19.gujarat.gov.in/>
wd:Q88973921	2020 COVID-19 pandemic in Manitoba	<https://www.gov.mb.ca/covid19/>
wd:Q87245450	2020 COVID-19 pandemic in Lebanon	<https://www.moph.gov.lb/>
wd:Q87245450	2020 COVID-19 pandemic in Lebanon	<https://corona.ministryinfo.gov.lb/>
wd:Q87245450	2020 COVID-19 pandemic in Lebanon	<https://www.the961.com/coronavirus/>
	2020 COVID-19 pandemic in	<https://coronavirusecuador.com/>

B

item	label	hashtag
wd:Q87705884	2020 COVID-19 pandemic in Kenya	COVID19KE
wd:Q87718451	2020 COVID-19 pandemic in Nigeria	CoronaVirusNigeria
wd:Q88622881	2020 COVID-19 pandemic in the European Union	CoronavirusEU
wd:Q88622881	2020 COVID-19 pandemic in the European Union	COVID19EU
wd:Q81068910	COVID-19 pandemic	COVID19FOAM
wd:Q86597695	COVID-19 pandemic in Brazil	covid19brasil
wd:Q87250732	2020 COVID-19 pandemic in Croatia	OstaniDoma
wd:Q87483673	2020 COVID-19 pandemic in Colombia	Covid19Colombia
wd:Q83873057	COVID-19 pandemic in Vietnam	CoronavirusVietnam
wd:Q83873387	2020 COVID-19 pandemic in Singapore	coronavirussingapore
wd:Q83872271	COVID-19 pandemic in mainland China	CoronaVirusChina
wd:Q83872271	COVID-19 pandemic in mainland China	coronaviruswuhan
wd:Q83872291	COVID-19 pandemic in Japan	CoronaVirusJapan
wd:Q83872398	2019–20 COVID-19 outbreak in South Korea	CoronaVirusSouthKorea
wd:Q83873548	2020 COVID-19 pandemic in Australia	coronavirusaus

Fig. S9. Online resource locations for information on COVID-19 regional outbreaks as of August 16, 2020 (Available at: https://w.wiki/cQo). A) Official websites (live data: https://w.wiki/bdt). B) Main hashtags (live data: https://w.wiki/bds)

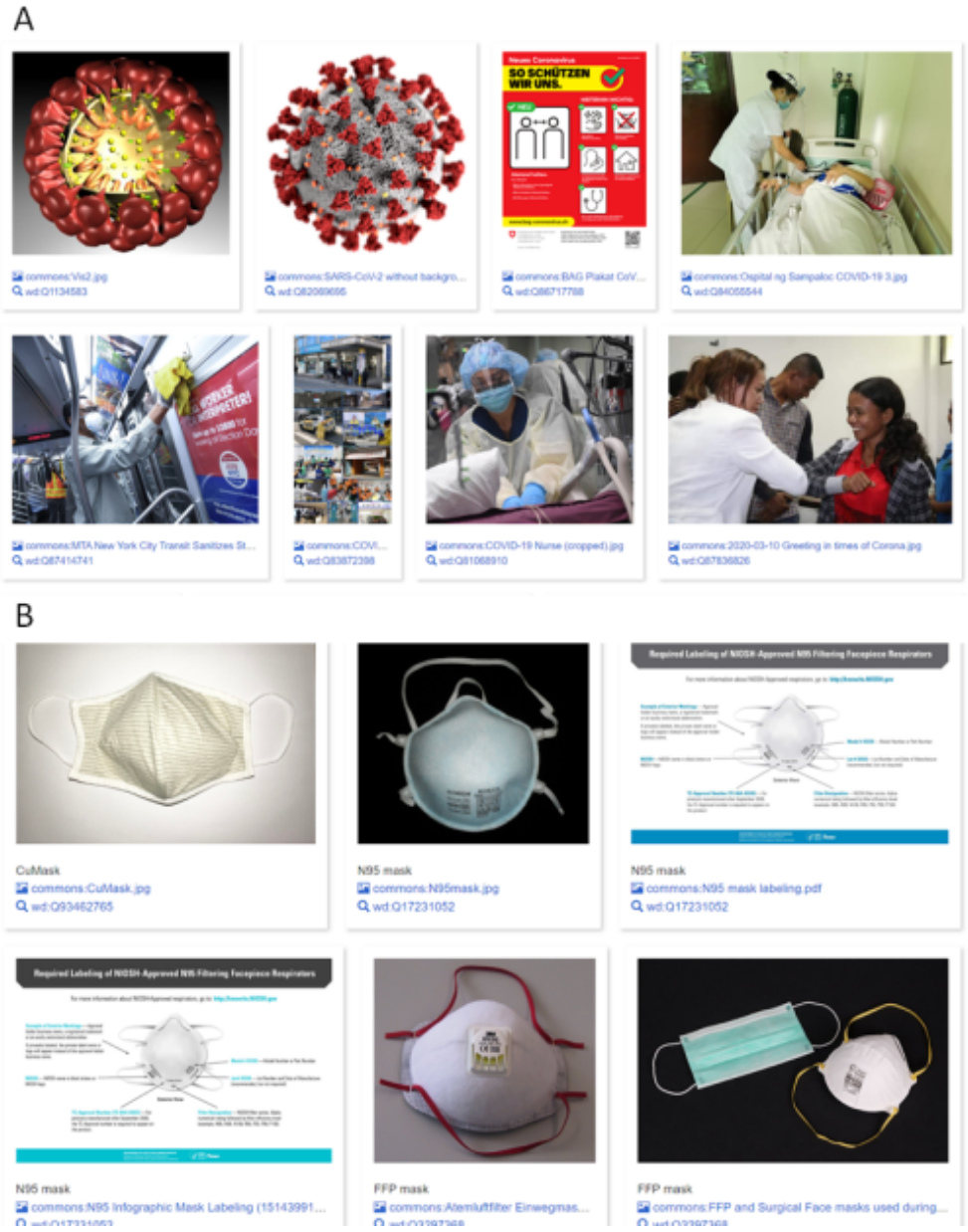

Fig. S10. COVID-related images based on structured data as of August 16, 2020 (Available at: https://w.wiki/cQt). Images in wikimedia commons used to be organised solely by a hierarchical category structure. Since 2019, structured data can be associated with images via Wikidata statements. A) Images from Wikimedia Commons about COVID-19 pandemic and SARS-CoV-2 with a CC-BY-compatible license (live data: https://w.wiki/Zsn). B) Images of face masks used during COVID-19 pandemic with a CC-BY-compatible license (live data: https://w.wiki/bzG).

Author responses to the comments

Article title: Representing COVID-19 information in collaborative knowledge graphs: the case of Wikidata

Tracking Number: 2572-3786

Authors: Turki et al.

Many thanks for the referee comments which have helped us improve the presentation and quality of the manuscript. We have responded to these point-by-point. Please find below our detailed response to the comments made by the reviewers. Adjusted parts are highlighted in yellow.

Editor comments

Comments	**Responses**
In light of the reviews I share the sentiment of one of the reviewers that the paper is better suited for a dataset description (i.e., a subgraph of Wikidata) rather than as a research paper.	We believe we are offering a research contribution, by showing on the basis of a specific dataset how collaborative editing and decentralized knowledge production can be particularly successful in addressing new, rapid phenomena. We also point out shortcomings of the model. This cannot be explained in the few pages that a dataset description allows for. For that reason, we believe that our paper fits the research criterion, even though we are happy that you find our contribution valuable as a dataset paper, too.
Contribution: Better define the novel contributions of the paper.	DONE We've reworked the abstract and introduction to better clarify and summarise what is new in the work presented. We have also highlighted how we are able to give a unique 'insider perspective' on the topic from being contributors to the development of the COVID-19 facets of Wikidata: "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, 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 Wikidata as a collaborative multidisciplinary large-scale knowledge graph can host COVID-19 data, integrate it with non-COVID-19 information and feed computer applications in an open and transparent way."
Coverage: Detail the exact coverage of the COVID data on Wikidata, e.g., are there data points missing such as economic data, what to do with conflicting/inconsistent information etc.?	The coverage of the COVID-19-related information is given with details in Section 3, "Visualizing facets of COVID-19 via SPARQL". The coverage includes not only epidemiological information but also biological and clinical aspects, research aspects and societal aspects for the pandemic including economic ones. The missing points for these aspects that are not efficiently covered in Wikidata are explained with examples in the same section. Concerning the management of conflicting and inconsistent COVID-19 information, this is a good point. Indeed, this is the topic of a second research paper just submitted for review to SWJ (available at https://doi.org/10.5281/zenodo.4008358, Tracking number: **2677-3891**). We thought it necessary to keep these two papers separate to avoid the length getting overwhelming. However, we added a paragraph at the end of the Data Model and Discussion sections to explain this: "Later, the developed semantic database for the pandemic is checked by multiple layers of validation. Methods include RDF triples defining

	conditions for the usage of Wikidata properties, RDF validation schemas built in Shape Expressions (ShEx) to verify the structural accuracy of the statement of an item included in a given Wikidata class, and logical constraints implemented in SPARQL to verify the consistency of relational and non-relational claims in Wikidata as well as several tools based on edit history of Wikidata such as ORES to identify and eliminate database vandalism. 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, the infrastructure developed for the validation of RDF data in Wikidata helps assure a high level of consistency of the Wikidata knowledge graph." "Greater consistency of structure and accuracy is partly due to the involvement of more contributors in Wikidata than in other open knowledge graphs. But it also stems from importing data from other rapidly-updated and curated databases (mainly from the linked open data cloud) and from verification by overlapping methods (e.g. ShEx schemas, SPARQL-based logical constraints and bot edits). The data validation infrastructure of Wikidata seems to be in accordance with the latest updates in knowledge graph evaluation and refinement techniques and explains in part the reasons behind the robustness of the data model of COVID-19 information in this open knowledge graph."
Statistical Rigor: Include statistical analysis to verify correlation analysis.	A very good point. For the observed correlations between content, editorship and readership in medical and COVID-19 content, we've added in all-vs-all statistical comparisons to rigorously back up the observations noted from tables 2-4. We have rephrased also the sentences for that claim to make it clear that the observations are descriptive in the phrase "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 10s for other sets described in Table 4." We agree that more in-depth tests, if planned a priori, could allow stronger causative claims. That is why, we have added correlation analysis of the statistical data about language representation: "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 10s 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 .65 to .97, median = .84, Supplementary Table S4), statistically significant even following stringent Bonferroni correction. To account for skew and data spanning multiple orders of magnitude, log10-transformed data was used for subsequent analyses. Pearson's correlation coefficients between all variables was high (Figure 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 .80 to .95. The smallest PCA loading and Spearman's correlation was for the number of viewers, which though still a strong association, was less correlated than the other variables by a substantial margin." "This is confirmed 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 and the maximum human development index for countries where that language is an official language. 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 similar positive skew to earlier data, so was also log10-transformed. Even though these analyses are necessarily exploratory, maximum development correlated more strongly than did the number of speakers (Figure 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.".
Usage: How can users access the information beyond the proposed SPARQL queries.	Good recommendation. We have now mentioned the web interface introduction: "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. 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. 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." We also expanded the final section on the SPARQL query service: "An important aspect of Wikidata's FAIRness is the Wikidata SPARQL query service (https://query.wikidata.org). 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. As shown here, SPARQL visualizations are an entrypoint for deeper insights into COVID-19, both regarding the biomedical facets of this still new disease, as well as into the societal details of the pandemic." Further, we added a note on data availability, making it clear that data on Wikidata can be downloaded via database dumps, the Wikibase API and API wrappers: "Another partner for FAIRness is user-friendly programmatic data access. Wikidata database dumps are available for download and local processing (https://www.wikidata.org/wiki/Wikidata:Database_download) in RDF, JSON and XML formats. 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/), arguably exposing the content even further."

Reviewer 1

Comments	Responses
Overall, the paper is written clearly, but it could be improved regarding the organisation to make its logical flow more manifest. The authors explain how Wikidata works briefly and illustrate the COVID related data in Wikidata. However, according to the paper's goal, the fitness of Wikidata for handling COVID related data, the following discussions would be insightful, i. COVID data specific issues, ii. different available methods and technologies for handling these issues and iii. the advantages of Wikidata's techniques for addressing the raised issues.	We have streamlined the organization of the sections (as outlined in the new section 1.2 "Organization of the manuscript") and the logical flow within and between them. The purpose of this paper is to highlight the value of Wikidata as a readily usable platform that can be quickly and flexibly customized to address emerging needs like visual representations of knowledge graphs pertaining to the COVID-19 pandemic. With regards to the suggested discussions, we have added a brief outline of COVID data specific issues to the introduction (cf. Section 1.1 "Data integration challenges") and refer to it when discussing related matters in various parts of the manuscript: "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[1] and naval[2] 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, systematically exposing the content of Wikidata to many eyes through its reuse in Wikipedia and SPARQL-based tools such

[1] https://en.wikipedia.org/wiki/COVID-19_pandemic_on_cruise_ships
[2] https://en.wikipedia.org/wiki/COVID-19_pandemic_on_naval_ships

	as Scholia and COVID dashboards, and using knowledge graph learning techniques to update items directly from textual databases like scholarly publications and electronic health records. Moreover, collaborative editing can cause several inaccuracies in the declaration of statements in open knowledge graphs disregarding the metadata standards of the knowledge bases. 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 and can consequently cause matters about the trustworthiness of the reuse of data. 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. Greater consistency of structure and accuracy is partly due to the involvement of more contributors in Wikidata than in other open knowledge graphs. But it also stems from importing data from other rapidly-updated and curated databases (mainly from the linked open data cloud) and from verification by overlapping methods (e.g. ShEx schemas, SPARQL-based logical constraints and bot edits). The data validation infrastructure of Wikidata seems to be in accordance with the latest updates in knowledge graph evaluation and refinement techniques and explains in part the reasons behind the robustness of the data model of COVID-19 information in this open knowledge graph." "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." "Concerning drugs, proteins, genes and taxons, Wikidata items are mainly assigned external identifiers in the major knowledge graphs for pharmacology (e.g. MassBank), for biodiversity (e.g. IRMNG), for genomics (e.g. Entrez Gene) and for proteomics (e.g. PDB) and are rarely linked to non-medical databases or to encyclopedias, as shown in Table 8." "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, 110323 COVID-19 full texts accessible on PubMed Central, 296450 COVID-19 publications on Dimensions, 211000 records on Semantic Scholar, 4778 records at ClinicalTrials.gov, 3295 records on arXiv ID, and 183 records on NIOSHTIC-2 as of February 17, 2021." "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 [27]) or multilinguality (e.g. how to handle concepts that are hard to translate), 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."
The paper avoids the discussion of underlying semantic technologies that are proposed and deployed for handling various aspects of complex	The purpose of the research paper was not to study the advantages of different semantic technologies to represent large-scale COVID-19

real data, including geospatial and time characteristics of data. For example, for quantifying a fact, there are different competing approaches, including property graph and RDF*. The explanation of Wikidata's quantifiers is not adequate regarding characterizing syntax and semantic of the deployed quantifying method and how Wikidata's way is more apt for modelling COVID data in comparison with the other methods.	information. The main aim of the paper is to demonstrate the usefulness of knowledge bases to handle and especially visualize COVID-19 data. Here, Wikidata's model offers a "good-enough" model to assess this statement. We agree that there might be other semantic technologies that are, rigorously, more adequate to represent specific bits of knowledge in general. However, in the context of COVID-19, RDF is a better choice than a property graph. In fact, property graphs are generally used in the context of social media, where the predicates of semantic relations do not matter as much. We further developed the Data Model section to explain this statement in details: "The advantage of RDF over other competing semantic data formats, particularly *property graph*, is that it applies reference schemas and consistency rules before assigning predicates to statements. Entries in RDF triple stores are predefined entities, rather than simple text strings, and structured into uni-directional statements [13]. 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 graph*, 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." We have also provided several other comparisons in favor of the Wikidata data model and the use of the RDF Format: "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, the infrastructure developed for the validation of RDF

	data in Wikidata helps assure a high level of consistency of the Wikidata knowledge graph." "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." "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. Wikidata is mainly enriched and adjusted by a community of over 25000 active human users on a daily basis 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." "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. In Wikidata, several initiatives such as WikiCite for scholarly information and Gene Wiki for genomic data 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." "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, the infrastructure developed for the validation of RDF data in Wikidata helps assure a high level of consistency of the Wikidata knowledge graph."

	"This process is called reification, and it is a common feature of many knowledge graphs such as DBpedia, Freebase, and YAGO. Although DBpedia and Freebase apply reification in a similar setting as in Wikidata, YAGO chooses to use N-Quads to represent the characteristics of a statement, implying that the additional feature is linked to the statement as a couple without the use of any predicate." "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. 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."
The authors do not provide convincing arguments to support how the characteristics of Wikidata addresses the specific issues that COVID-19 related data raised.	Our main argument is that Wikidata's versatility (and the community-centric approach) are particularly relevant for addressing rapid and emerging phenomena, such as COVID-19 pandemic. As mentioned in response to your first comment, we also added an outline of COVID-specific data issues and comment on how Wikidata addresses them. We additionally focus on showcasing a snapshot of how the COVID-19 knowledge graph of Wikidata can be used to support computer applications, particularly the SPARQL-based visualization of multidisciplinary information about COVID-19. We have added a paragraph in the conclusions to clarify this for systematic knowledge representation.intent: "We have shown how the community-driven and not centrally coordinated approach to editing 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."

Arguing a Knowledge Base (e.g. Wikidata) is a reasonable solution for handling COVID-19 related data is an exciting idea. However, the authors do not provide convincing arguments to support how the characteristics of Wikidata addresses the specific issues that COVID-19 related data raised.	As detailed in response to your first comment, we have added a brief outline of COVID data specific issues to the introduction (cf. Section 1.1 "Data integration challenges") and highlighted more clearly (particularly in Section 2 "Wikidata as a semantic resource for COVID-19") how Wikidata addresses these challenges.

Reviewer 2

Comments	**Responses**
In Introduction, the authors talk about the benefit and drawback of the 'community developed ontology and typology' (second paragraph). In terms of the drawback, it claims that "it makes methodical planning of the whole structure and its granularity very difficult". However, in the main text I do not clearly see how these issues are addressed in this project.	We have expanded the sentence to clarify that we specifically mean the pros and cons of a lack of a centralized coordination: "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 [10]: 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." We have also expanded this part to explain how these issues are addressed in this project: 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, and its existing community has been using it to capture COVID-19-related knowledge right from the start. 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. Third, both the Wikidata platform and the Wikidata community are highly multifaceted, multilingual and multidisciplinary. 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, 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 or through overlap with the Wikipedia community that has been covering disasters for two decades.
Data Model section: The authors claim that '... an ontological database representing all aspects of the outbreak'. Is it really the case? For example, does it cover economic aspects that include information about the unemployment rate and supply chain disruption during this outbreak? I think it is a too ambitious statement.	Wikidata represents many facets of the COVID-19 pandemic and the cited examples can be represented too. Indeed, a Wikidata property for unemployment rate already exists (P1198). However, the representation of these facts using SPARQL queries is limited by the lack of volunteers enriching socio-economic information of countries in Wikidata. We adjusted the claim to be "an ontological database representing many aspects of the SARS-CoV-2 outbreak". An example of the limitation of Wikidata for assessing societal aspects of the COVID-19 pandemic is added to Visualizing facets of COVID-19 via SPARQL section (Societal aspects): "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."
Data Model section: What exact lessons are learned from the Zika pandemic?	DONE We added several lines about the lessons learned from the Zika pandemic throughout the manuscript: "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, 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 or through overlap with the Wikipedia community that has been covering disasters for two decades." "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 and led to the formation of WikiProject Zika Corpus. This Zika project—itself inspired by dedicated Wikiprojects for Medicine and for Source Metadata— 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."

Data Model section: The authors mention '… could all be represented in Wikidata if matters related to the coverage and conflicts of information from multiple sources are solved'. In fact, it would be great if the authors can discuss about how does the model solve the issue about conflicting statements in the project? In Covid-19, it becomes particularly essential as we see various reported 'facts' that are conflicting/inconsistent with each other. In addition, what does 'coverage' mean here? Spatial coverage? Temporal coverage? Or property coverage? A little bit confusing.	The coverage of the COVID-19-related information is given with details in "Visualizing facets of COVID-19 via SPARQL". The coverage includes not only spatial information but also temporal and social information for the pandemic including economic ones. The missing points for these aspects that are not efficiently covered in Wikidata are explained with examples in the same section. The statement in the Data Model section has been adjusted: "… 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". Concerning the management of conflicting and inconsistent COVID-19 information, it will be the topic of a second research paper sent for review to SWJ (available at https://doi.org/10.5281/zenodo.4008358, Tracking number: **2677-3891**). It would be overwhelming to explain this in this research paper. However, we added a paragraph at the Data Model and Discussion sections to explain this: "Later, the developed semantic database for the pandemic is checked by multiple layers of validation. Methods include RDF triples defining conditions for the usage of Wikidata properties, RDF validation schemas built in Shape Expressions (ShEx) to verify the structural accuracy of the statement of an item included in a given Wikidata class, and logical constraints implemented in SPARQL to verify the consistency of relational and non-relational claims in Wikidata as well as several tools based on edit history of Wikidata such as ORES to identify and eliminate database vandalism. 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, the infrastructure developed for the validation of RDF data in Wikidata helps assure a high level of consistency of the Wikidata knowledge graph." "Greater consistency of structure and accuracy is partly due to the involvement of more contributors in Wikidata than in other open

	knowledge graphs. But it also stems from importing data from other rapidly-updated and curated databases (mainly from the linked open data cloud) and from verification by overlapping methods (e.g. ShEx schemas, SPARQL-based logical constraints and bot edits). The data validation infrastructure of Wikidata seems to be in accordance with the latest updates in knowledge graph evaluation and refinement techniques and explains in part the reasons behind the robustness of the data model of COVID-19 information in this open knowledge graph."
Language Representation section: Figure 4E is confusing, the x-axis is the rank of languages based on their usages? What does y-axis mean then? The sentence: "The degree of translation of that information is increasingly high with an important representation of the concepts in more than 50 languages (Figure 4E)" does not help to understand the figure.	We have reformulated the inline description of Figure 4E to reflect its outcomes: "The degree of translation of that information is interestingly high with an important representation of the concepts in more than 50 languages (Fig. 4E). In fact, 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." We have adjusted the title of the Figure to clarify: "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" We have also added titles to the x-axis and y-axis of the Figure to improve its understandability: x-axis: Rank of language (per each category A-D) y-axis: Coverage of concepts
Language Representation section: More importantly, there are multiple correlation analyses in this section. However, no statistical analysis is applied at all. The conclusions are all made by arbitrarily checking the tables. For example, the statement "Despite several	Well noticed, thank you for pointing that out. We have rephrased the sentences for that claim to make it clear that the observations are descriptive in the phrase "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

differences like the higher visibility of Asian language... the query results largely match the literature-derived data ...” has to be justified in a more scientific way, e.g., by statistical testing.	languages in the Wikidata COVID-19 subset, while being absent from the top 10s for other sets described in Table 4.” We agree that more in-depth tests, if planned a priori, could allow stronger causative claims. That is why, we have added correlation analysis of the statistical data about language representation: “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 10s 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 .65 to .97, median = .84, Supplementary Table S4), statistically significant even following stringent Bonferroni correction. To account for skew and data spanning multiple orders of magnitude, log10-transformed data was used for subsequent analyses. Pearson’s correlation coefficients between all variables was high (Figure 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 .80 to .95. The smallest PCA loading and Spearman’s correlation was for the number of viewers, which though still a strong association, was less correlated than the other variables by a substantial margin.” “This is confirmed 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 and the maximum human development index for countries where that language is an official language. 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 similar positive skew to earlier data, so was also log10-transformed. Even though these analyses are necessarily exploratory, maximum development correlated more strongly than did the number of speakers (Figure 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.". We have expanded our treatment of the language representation in the following ways: We present more descriptive information, we have switched to nonparametric statistical analysis to better model the data distributions, we note that for the core set of variables, the effect sizes are all large enough to survive even stringent post hoc correction to control

	family-wise type I error rate. We added the number of speakers and maximum economic development as additional variables for supplemental analyses. These were available for fewer languages, so here we emphasize differences in effect size (Cohen's *q*) rather than significance testing. The results indicate that maximum economic development is substantially more related than the number of speakers to the medical and Wikidata metrics, but with negligible differences in association in number of page views.
Database Alignment section: This section lists multiple alignment tables for different domains. However, how are these alignments accomplished? Any automated algorithms are used or totally based on human efforts? Have these alignments been evaluated?	Excellent point. As Wikidata's graph is collaborative, any effort of database alignment does include some parts of manual, human efforts. Some of the work has been via reconciling of databases and semi-automatic triple adding via tools such as https://github.com/SuLab/WikidataIntegrator or https://quickstatements.toolforge.org/#/. A match based on identifiers like DOI or PubMed ID is usually enough for a reliable key to reconcile to Wikidata. It would be great to evaluate these alignments. It is hard, however, to devise an automatic way, as there is no gold standard. The alignments generally follow an "Anyone can say Anything about Anything" assumption. We added a paragraph to the database alignment section to try to clarify it: "The alignment of Wikidata entities to other entries on different databases is a collaborative process which, as 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 for locating the database ID. 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.".

	We also added several lines to data model section in this particular context: "Wikidata items are assigned their identifiers in external databases, including semantic resources, using human efforts and tools such as Mix'n'match. 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 and federation in the context of the linked open data cloud."
Visualizing facets of COVID-19 via SPARQL and Conclusion: It is great to see the authors bring up a relative comprehensive and well organized list of SPARQL queries, and demonstrated several promising visualization in the paper. However, I am wondering how accessible and easy for a non-SPARQL expert to explore the graph (or simply understand the query)? Do the authors have any empirical examples/cases to show how useful the graph has been to domain experts/general public? In Table S2, it seems to be a list about fulfilled tasks; but I do not find more contexts related to this table. Maybe use one of the rows in this table as an example to elaborate would help readers understand the value of the proposed graph.	We do not have an analysis of popularity of SPARQL use by the general public. However, we would like to note that - typically for the open source movement - most beginners find copying ready-made examples and paragon syntax useful, and Wikidata provides plenty of examples, which are easily modifiable. One of the advantages of this approach is that users do not have to have expert understanding of SPARQL to be able to slightly modify the code to reach satisfactory results. We added a note on that in the discussion: "One of the features of Wikidata is also providing hundreds of exemplary SPARQL queries, 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. As a result, Wikidata users do not have to be SPARQL experts to arrive at results that are useful to them." With regards to Table S2, we have clarified both its purpose and content: "Sample SPARQL queries for data visualizations commonly included in Wikidata-based COVID-19 dashboards are available at Supplementary Table S2 to show the variety of visualizations that can be generated using the Wikidata Query Service 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)." "List of sample queries on COVID-19. The information contained therein is similar to visualizations in many stand-alone COVID-19 dashboards, covering an overview of COVID-19, international situation, international daily epidemiological evolution, Tunisian daily epidemiological evolution, Tunisian governorate-level situation, Tunisian correlations, and worldwide correlations. Each of the sheets has a Title column with a brief summary for each query and a URL column with a link to the live record on Wikidata."
Last but not least, the authors have to proofread the paper substantially. There are many long sentences, inconsistent uses of terms, typos, duplicates, and many weird sentences. In general, the paper is not that easy to follow. For example, solely in the first paragraph of Section 5.2: a). whereas others common visualization --> other b). from scratch from granularity --> one 'from' has to be deleted c). its change over time over time --> duplicates d). Wikidata's granularity and collaborating … --> What does 'wikidata's granularity' mean here?	DONE We have gone through the manuscript and brushed its grammar. As for the definition of granularity, it is the representation of the COVID-19 information at a narrow and specific scale such as the famous COVID-19 mortality and morbidity cases. We have adjusted the manuscript to make this very clear: "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."
Page 2: a). basing --> based b). entities named items --> entities, named items	DONE Adjusted
Page 3:	DONE Adjusted

>17,000 (what is this number? Cases? Deaths?)	
Page 5: Table S1 --> Table 1	DONE Adjusted
page 13: table S2 --> Table S2	DONE Adjusted
page 14: a). allowed --> allows b). WIkidata --> Wikidata	DONE Adjusted

Reviewer 3

Comments	**Responses**
Some features the authors discussed about Wikidata are in fact well-known. For example, the data model, the multilingual features as well as its alignment to other databases. Since this paper is explicitly about the COVID-19 efforts of Wikidata. I suggest the authors highlight the specific features Wikidata considers for COVID-19.	It is a fair comment - Wikidata's systemic features have certainly been discussed before. However, the interactions between these features and the user community in a global disaster response context have not been discussed in detail before, and given the diversity of topics covered by SWJ, a longer introduction could make it easier for some readers to understand the work in its sociotechnical context. We also streamlined the text such that it conveys more clearly why Wikidata is so well suited for addressing rapid, emerging phenomena such as the COVID-19 pandemic. Even though we focus on COVID-19 efforts, we believe that our conclusions reach beyond that - yet, for the reader, it is crucial to understand the background. For that reason, we would like to keep the remaining descriptions of relatively well-known Wikidata, even though we understand that for some readers, it will be repetitive (but for some others, it will provide a crucial introduction to the discussed topic).

	With regards to the specific Wikidata features relevant for COVID-19 and as mentioned in response to a similar comment by Reviewer 1, we have added a brief outline of COVID data specific issues to the introduction (cf. Section 1.1 "Data integration challenges") and highlighted more clearly (particularly in Section 2 "Wikidata as a semantic resource for COVID-19") how Wikidata addresses these challenges. "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[3] and naval[4] 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, systematically exposing the content of Wikidata to many eyes through its reuse in Wikipedia and SPARQL-based tools such as Scholia and COVID dashboards, and using knowledge graph learning techniques to update items directly from textual databases like scholarly publications and electronic health records. Moreover, collaborative editing can cause several inaccuracies in the declaration of statements in open knowledge graphs disregarding the metadata standards of the knowledge bases. 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 and can consequently cause matters about the trustworthiness of the reuse of

[3] https://en.wikipedia.org/wiki/COVID-19_pandemic_on_cruise_ships
[4] https://en.wikipedia.org/wiki/COVID-19_pandemic_on_naval_ships

	data. 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. Greater consistency of structure and accuracy is partly due to the involvement of more contributors in Wikidata than in other open knowledge graphs. But it also stems from importing data from other rapidly-updated and curated databases (mainly from the linked open data cloud) and from verification by overlapping methods (e.g. ShEx schemas, SPARQL-based logical constraints and bot edits). The data validation infrastructure of Wikidata seems to be in accordance with the latest updates in knowledge graph evaluation and refinement techniques and explains in part the reasons behind the robustness of the data model of COVID-19 information in this open knowledge graph." "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." "Concerning drugs, proteins, genes and taxons, Wikidata items are mainly assigned external identifiers in the major knowledge graphs for pharmacology (e.g. MassBank), for biodiversity (e.g. IRMNG), for genomics (e.g. Entrez Gene) and for proteomics (e.g. PDB) and are rarely linked to non-medical databases or to encyclopedias, as shown in Table 8."

	"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, 110323 COVID-19 full texts accessible on PubMed Central, 296450 COVID-19 publications on Dimensions, 211000 records on Semantic Scholar, 4778 records at ClinicalTrials.gov, 3295 records on arXiv ID, and 183 records on NIOSHTIC-2 as of February 17, 2021." "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 [27]) or multilinguality (e.g. how to handle concepts that are hard to translate), 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."
The contribution of this paper is not clear enough to me in the beginning. In the end, I realize the authors are responsible for managing COVID-19 information in Wikidata. I suggest the author list the contribution at the beginning of this paper.	DONE We added our contribution to the development of Wikidata's COVID-19: "In this research paper, we report on the efforts of the Wikidata community (including our own) to meet these challenges by serving 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, 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 Wikidata as a collaborative multidisciplinary large-scale knowledge graph can host COVID-19 data, integrate it with non-COVID-19 information and feed computer applications in an open and transparent way.”
The author claims this paper is a research paper while I think this is a dataset paper. I do think dataset papers are also very important, especially for the Semantic Web community. So please rethink the paper type you want to submit here.	We believe we are offering a research contribution, by showing on a specific dataset how collaborative editing and decentralized knowledge production can be particularly successful in addressing new, rapid phenomena. We also point out the shortcomings of the model. For that reason, we believe that our paper fits the research criterion, even though we are happy that you find our contribution valuable as a dataset paper, too.

Thank you again for helping us to improve the quality and presentation of this manuscript.