Representing COVID-19 information in collaborative knowledge graphs: a study of Wikidata

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

[a] Faculty of Medicine of Sfax, University of Sfax, Sfax, Tunisia
[a] turkiabdelwaheb@hotmail.fr
[b,f] Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia
[b] mohamedali.hadjtaieb@gmail.com, [f] mohamed.benaouicha@fss.usf.tn
[c] La Trobe University, Melbourne, Victoria, Australia
[c] t.shafee@latrobe.edu.au
[d] Computational Systems Biology Laboratory, University of São Paulo, São Paulo, Brazil
[d] tiago.lubiana.alves@usp.br
[e] Department of Management in Networked and Digital Societies, Kozminski University, Warsaw, Poland
[e] darekj@kozminski.edu.pl
[g] Web Semantics Oviedo (WESO) Research Group, University of Oviedo, Spain
[g] jelabra@gmail.com
[h] Faculty of Medicine, Hashemite University, Zarqa, Jordan
[h] mossab748@gmail.com
[i] Institute of Child Health (ICH), Kolkata, India
[i] Medica Superspecialty Hospital, Kolkata, India
[i] das.diptanshu@gmail.com
[j] School of Data Science, University of Virginia, Charlottesville, Virginia, United States of America
[j] dm7gn@virginia.edu

[k] Project members: Jan Ainali, Susanna Ånäs, Erica Azzellini, Mus'ab Banat, Mohamed Ben Aouicha, Diptanshu Das, Lena Denis, Rich Farmbrough, Daniel Fernández-Álvarez, Konrad Foerstner, Jose Emilio Labra Gayo, Mohamed Ali Hadj Taieb, James Hare, Alejandro González Hevia, David Hicks, 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, Bodhisattwa Mandal, Daniel Mietchen, Nandana Mihindukulasooriya, Mahir Morshed, Peter Murray-Rust, 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

***Corresponding author:** Daniel Mietchen
School of Data Science, University of Virginia, Charlottesville, Virginia, United States of America
dm7gn@virginia.edu

Abstract: Information related to the COVID-19 pandemic ranges from biological to bibliographic and from geographical to genetic. Wikidata is a vast interdisciplinary, multilingual, open collaborative knowledge base of more than 88 million entities connected by well over a billion relationships and is consequently a web-scale platform for broader computer-supported cooperative work and linked open data. Here, we introduce four aspects of Wikidata that make it an ideal knowledge base for information on the COVID-19 pandemic: its flexible data model, its multilingual features, its alignment to multiple external databases, and its multidisciplinary organization. The structure of the raw data is highly complex, so converting it to meaningful insight requires extraction and visualization, the global crowdsourcing of which adds both additional challenges and opportunities. The created knowledge graph for COVID-19 in Wikidata can be visualized, explored and analyzed in near real time by specialists, automated tools and the public, for decision support as well as

educational and scholarly research purposes via SPARQL, a semantic query language used to retrieve and process information from databases saved in Resource Description Framework (RDF) format.

Keywords: Public health surveillance, Wikidata, Knowledge graph, COVID-19, SPARQL, Community curation, FAIR data, Linked Open Data

1. Introduction

The COVID-19 pandemic is complex and multifaceted and touches on almost every area of current life. Coordinating efforts to systematize and formalize knowledge about COVID-19 in a computable form is key in accelerating our response to the pathogen (Domingo-Fernández, et al., 2020). There are already attempts at community-based ontologies of COVID-19 knowledge and data (He, et al., 2020), as well as efforts to aggregate expert data (Ostaszewski, et al., 2020). Many open data initiatives are started spontaneously (Desvars-Larrive, et al., 2020; Liu, et al., 2020; Wang, et al., 2020). However, we believe that more attention in the research world is needed for Wikidata, in particular as a platform for COVID-19 knowledge graph visual representation. The interconnected, interdisciplinary, and international nature of the pandemic makes Wikidata a well-suited knowledge base to collate and make sense of this information (Turki, et al., 2019).

Wikidata is a large-scale, collaborative, open-licensed, multilingual knowledge base that is both human- and machine-readable. Notably, it is available in the standardised RDF (Resource Description Framework) format, where data is organised into entities named *items* and the relationships that connect them, named *properties* (Vrandečić & Krötzsch, 2014). Wikidata is a peer production project, developed under the umbrella of the Wikimedia Foundation. Similarly to Wikipedia, it relies on collaborative development design and is both a-hierarchical and largely uncoordinated (Jemielniak, 2014). As a result, it develops entirely organically, basing on the editor community's consensus. This community develops ontologies and typologies used in the database. This spontaneous 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 (Konieczny, 2010). On the other, it provides for an unparalleled flexibility and versatility of uses.

One of Wikidata's key strengths is that each item can be understood by both machines and humans as it represents data in the form of triples (Vrandečić & Krötzsch, 2014). 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 (Kirk, 2016). The main way of doing this is by visualising some subset of the data: effectively, 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[1], a semantic query language dynamically extracting triple information from large-scale knowledge graphs. Here, we present how various types of data

[1] the recursive acronym for "SPARQL Protocol and RDF Query Language", the current version of which is SPARQL 1.1.

related to the COVID-19 pandemic are currently represented in Wikidata thanks to the flexible structure of the database and how useful visualisations for different subsets of the data linked to COVID-19 within the Wikidata knowledge base can be generated. In this way, we can understand facets of how the machine world 'sees' Wikidata's COVID-19 knowledge.

In this research paper, we describe the data model of Wikidata in general and in the context of COVID-19 pandemic (Section 2). Then, we give an overview of the language support (Section 3) and database alignment (Section 4) 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 5). Finally, we discuss the outcomes of the open development of the COVID-19 knowledge graph in Wikidata (Section 6) and we draw conclusions and future directions for this research paper (Section 7).

2. Data model

In Wikidata, each concept has an item (a human, disease, drug, city, etc), 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 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], and *part of* [P361]) or non-taxonomic (e.g. *significant person* [P3342], *drug used for treatment* [P2176], *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]). 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 (Turki, et al., 2019). This comes together to create an integrated network of 88 million items interlinked by over a billion statements.

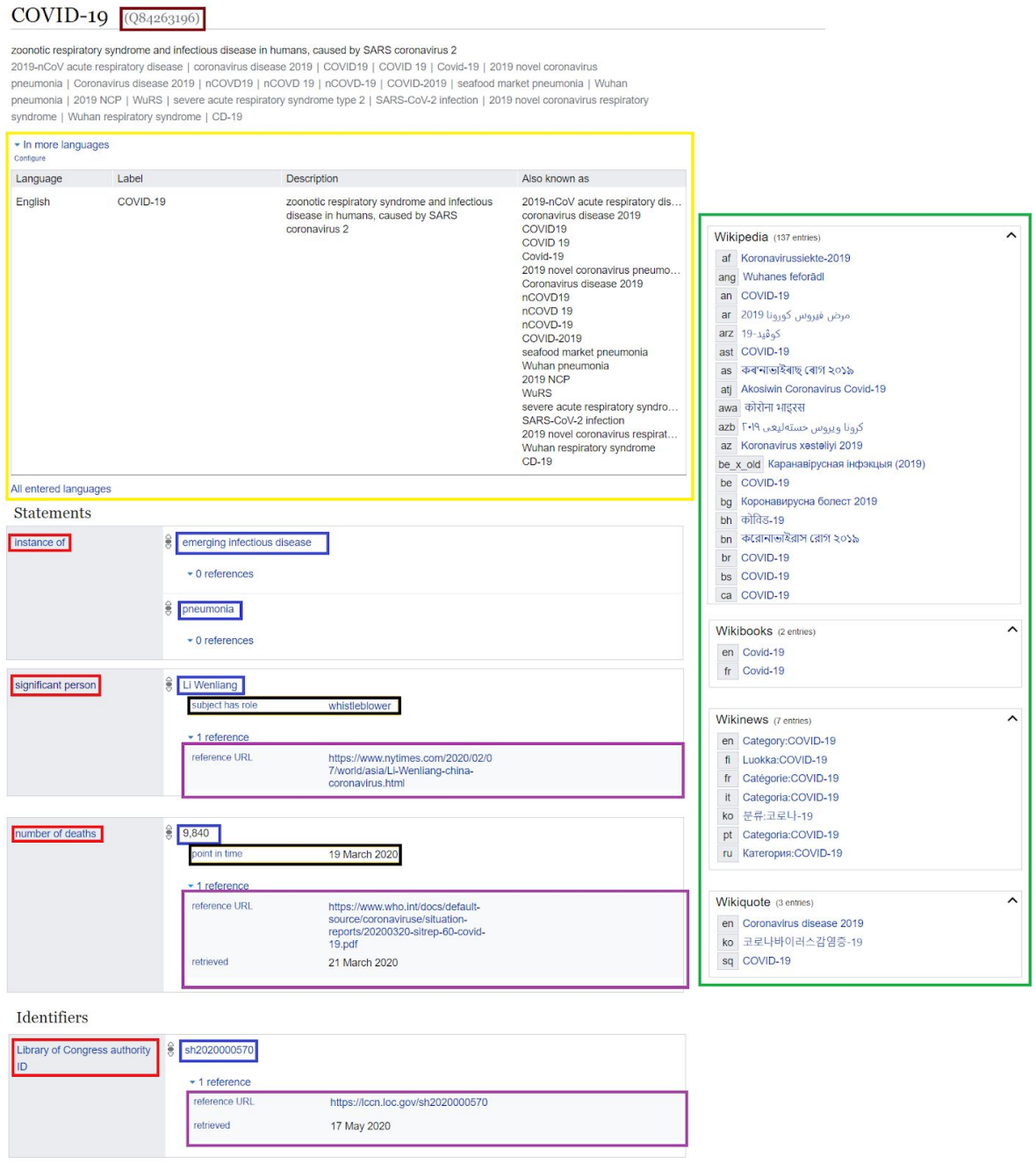


Figure 1: Data Structure of a Wikidata item. The simple, consistent structure of a Wikidata item makes it both human- and machine-readable. Each Wikidata item has a unique identifier (Brown). They can have labels, descriptions and aliases in multiple languages (Yellow). They then include any number of statements having predicates (Red), objects (Blue), qualifiers (Black) and references (Purple) where the subject is the item. Finally, where additional Wikimedia resources are available about an item's topic, those are listed (Green). Source: https://www.wikidata.org/wiki/Q84263196, available at: https://w.wiki/auF.

In the context of the COVID-19 pandemic, and building on pilot work that was started at the onset of the Zika pandemic (Ekins et al., 2016), an ontological database representing all the aspects of the outbreak has been represented in Wikidata. There are three main items that form the core of this structure (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 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[2], launched in March 2020 (Waagmeester, et al., 2020a). Indeed, WikiProject COVID-19 itself has a Wikidata item (Q87748614) and items are linked to this using the property *on focus list of Wikimedia project* (P5008). This WikiProject has built heavily on the lessons learnt from the pilot work during the Zika pandemic (Ekins et al., 2016).

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 high-scale knowledge graph for COVID-19, where the main Wikidata items are *COVID-19 pandemic* (Q81068910), *SARS-CoV-2* (Q82069695) and *COVID-19* (Q84263196) and where the mainly developed Wikidata classes are *disease outbreaks* (Q3241045) in regions such as continents, sovereign states, and constituent states, *COVID-19 tracing apps* (Q89288125), *vaccine candidates* (Q28051899), *scholarly articles* (Q13442814) and *COVID-19 dashboards* (Q90790055). This graph is augmented by biomedical, geographical and other information already available in Wikidata and representing an important overview of clinical knowledge (Turki, et al., 2019; Waagmeester, et al., 2020a). This action allows the inclusion of other classes to COVID-19 knowledge graphs, including *genes* (Q7187), *proteins* (Q8054), and *biological processes* (Q2996394) as well as the definition of semantic relations between COVID-19-related items and other Wikidata items. This, consequently, allows the expansion of the coverage of COVID-19 information in Wikidata and a better characterization of COVID-19-related items.

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

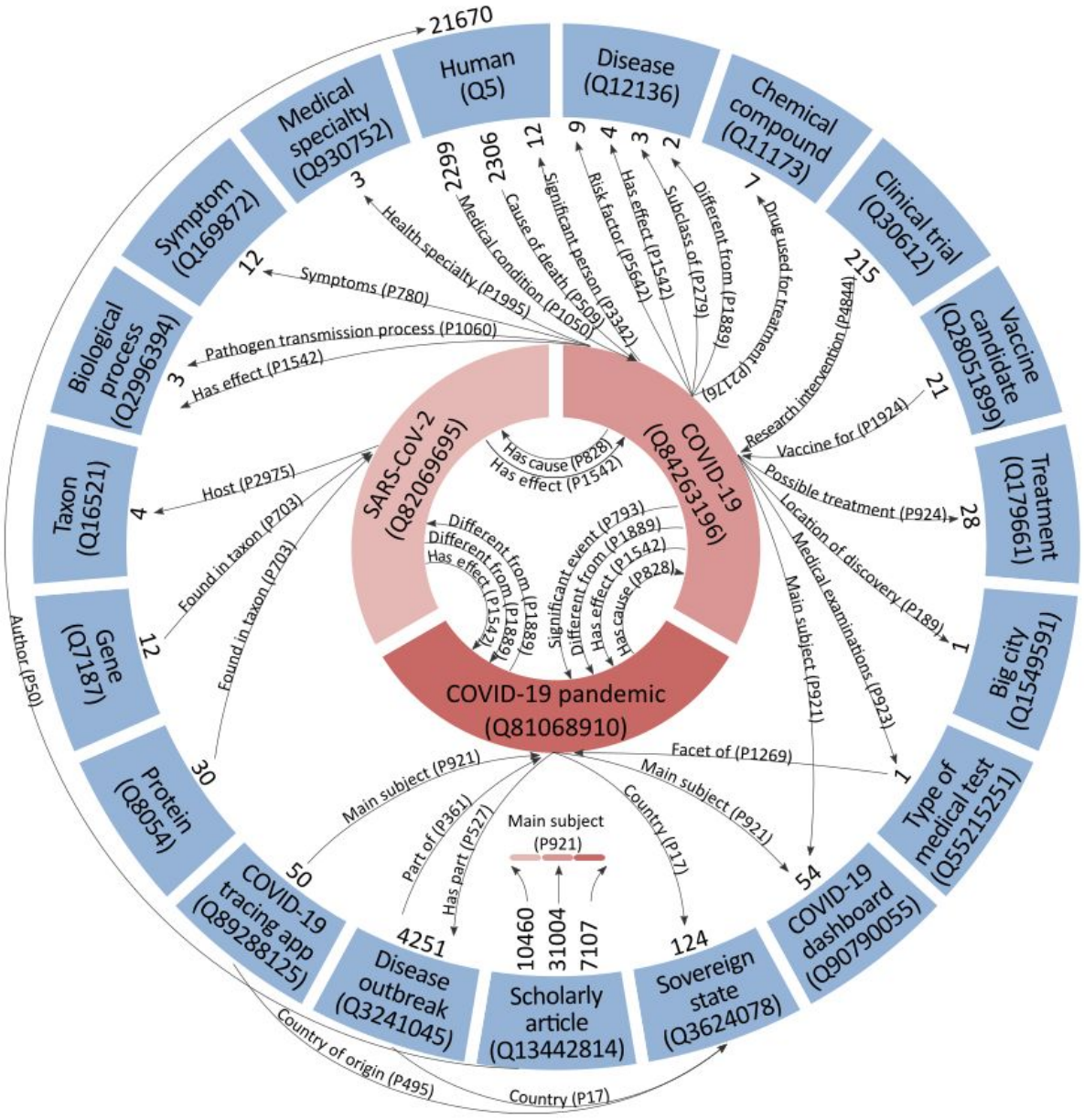


Figure 2: Simplified skeleton of the data model of COVID-19 information on Wikidata. The three main COVID-related items (the 'C3 items')[3] are represented in red, selected classes of items related to these are shown in blue, with the relations between them represented as arrows. The number of statements 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[4], available at: https://w.wiki/auD , license: CC BY 4.0.

Given the distinctive feature of knowledge graphs that assigns triples to items where the object is not an item (Ehrlinger & Wöß, 2016), Wikidata items are assigned an important number of non-relational statements for various purposes. On the one hand, items are assigned their identifiers in external databases, including semantic resources, using human efforts and tools such as Mix'n'match (Malyshev, et al., 2018). These links make Wikidata a

[3] COVID and C3 stand for any subset of {*COVID-19* [Q84263196], *SARS-CoV-2* [Q82069695], *COVID-19 pandemic* [Q81068910]}.

[4] 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.

key part 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 (Burgstaller-Muehlbacher, et al., 2016) and federation (Malyshev, et al., 2018) in the context of the linked open data cloud (Debattista, et al., 2018). 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.

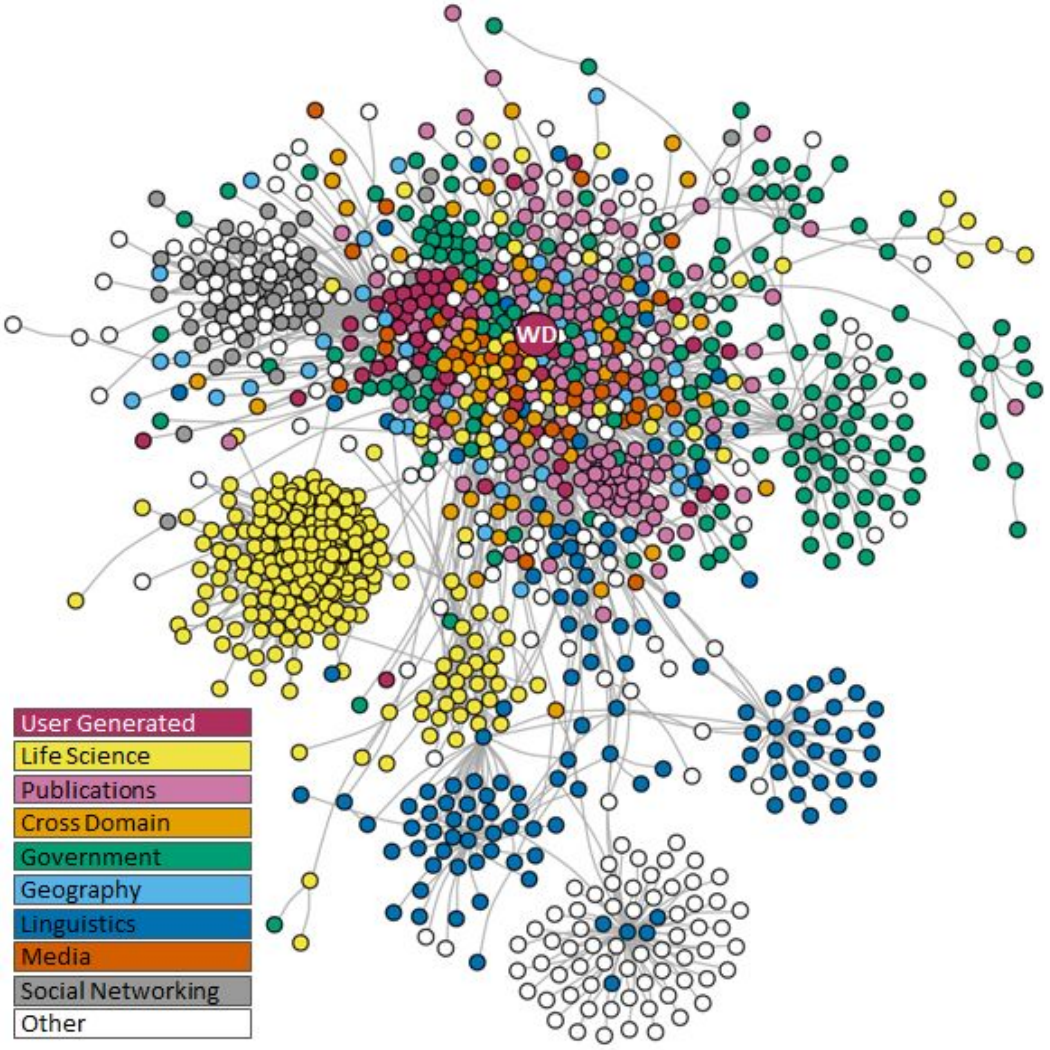


Figure 3: Wikidata in the Linked Open Data Cloud. Databases indicated as circles (with Wikidata indicated as 'WD'), with grey lines linking databases in the network if their data is aligned, as of 20 August 2020 (available at: https://w.wiki/bYM, license: CC BY 4.0).

On the other hand, disease outbreak items for the COVID-19 pandemic are assigned statistical statements to outline the evolution of the epidemiological status of different countries. The properties used to define these 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. Some of the 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[5] 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 coverage and conflicts of information from multiple sources are solved), the following measures can be inferred:

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

This set of COVID-19 information is integrated into Wikidata using human efforts, the QuickStatements tool[6], the Wikidata API[7], and bots mainly written in Python (e.g. CovidDatahubBot[8]), which explains its quantity and coverage (Turki, et al., 2019).

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

Wikidata ID	Name	Description
Properties for the alignment with scholarly databases		
P496	ORCID iD	identifier for a researcher (Open Researcher and Contributor ID)
P1153	Scopus Author ID	identifier for an author in the Scopus bibliographic database
P214	VIAF ID	identifier for the Virtual International Authority File database
P7859	WorldCat Identities ID	entity on WorldCat for authority control of authors' data
P1053	ResearcherID	identifier for a researcher in a system for scientific authors, primarily used in Web of Science
Properties for the alignment with clinical language resources and encyclopedias		
P494	ICD-10	identifier in the ICD catalogue codes for diseases - Version 10
P672	MeSH tree code	Medical Subject Headings (MeSH) codes are an index and thesaurus for the life sciences (≠ MeSH ID, P486)
P1417	Encyclopædia Britannica Online ID	identifier for an article in the online version of Encyclopædia

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

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

[7] 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

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

		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

3. Language 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[9], BabelMeSH[10], and ICD-10[11] (Turki, et al., 2019). 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

[9] http://www.snomed.org/snomed-ct/sct-worldwide (Accessed on November 11, 2019): SNOMED-CT supports English, French, Danish, Dutch, Spanish, Swedish, and Lithuanian.

[10] https://lhncbc.nlm.nih.gov/project/babelmesh-and-pico-linguist (Accessed on November 11, 2019): BabelMeSH supports Arabic, Chinese, Dutch, English, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish, and Swedish.

[11] ICD-10: International Classification of Diseases, 10th Revision (Jetté, et al., 2010): 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.

an important representation of the concepts in more than 50 languages (Fig. 4E). 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 (Liu, Fontelo, & Ackerman, 2006; Henriksson, Skeppstedt, Kvist, Duneld, & Conway, 2013) and is in line with efforts of generating multilingual language resources to be used for natural language processing purposes in clinical contexts (De Melo & Weikum, 2010).

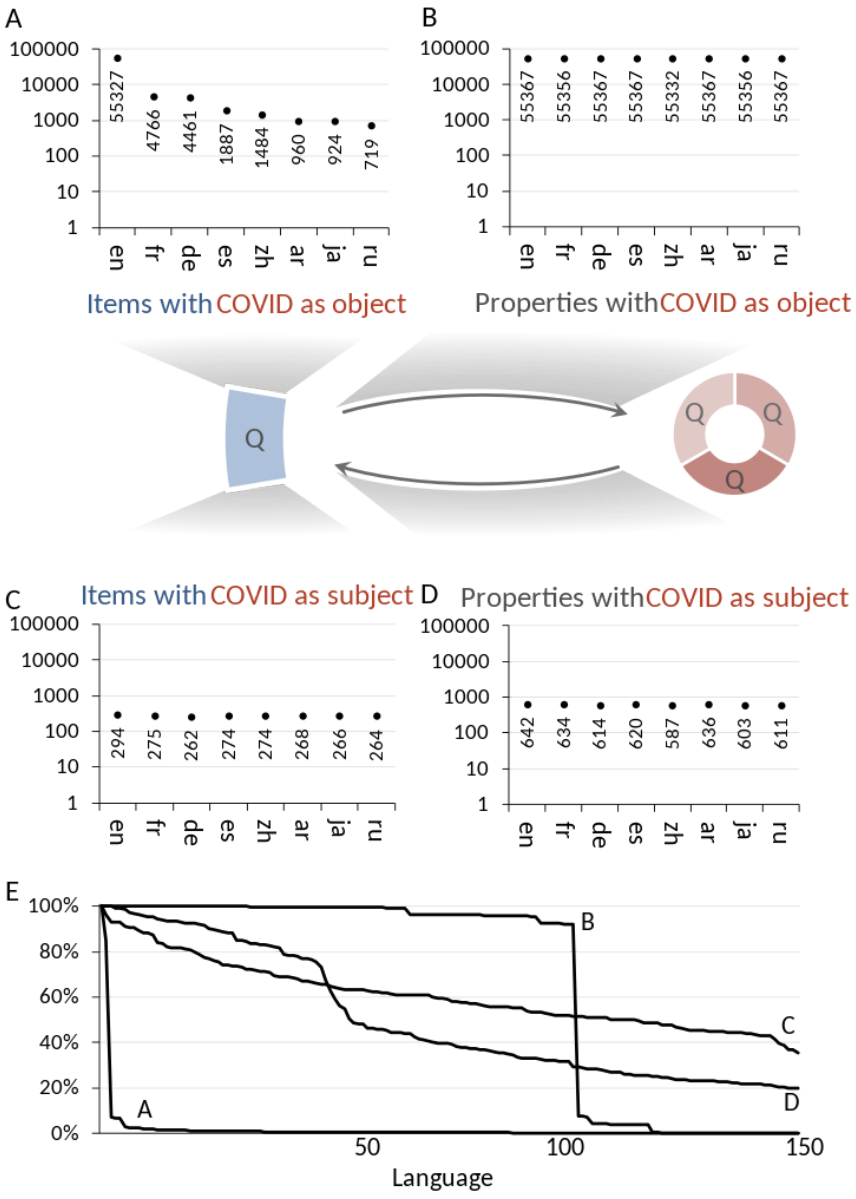


Figure 4: Language representation of COVID-19-related statements. A-D) Language coverage for items and properties used in statements when either the object or subject is one of the three COVID-related items (as per figure 2; note: log y-axis). The eight most common languages in Wikidata are shown: en=English, fr=French, de=German, es=Spanish, zh=Chinese, ar=Arabic, ja=Japanese, ru=Russian. E) Extended data for panels A-D, with the top 150 languages ordered by usage for each item (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/YjŚ, https://w.wiki/Yk3, https://w.wiki/Yk5, and https://w.wiki/Yk6)

The better coverage of English is explained in part by the higher support of this language in both biomedical language resources (Freitas, Schulz, & Moraes, 2009) and Wikipedia (Shafee, et al., 2017). Cooperation with publishers such as Cochrane has a significant effect on English Wikipedia coverage, too (Jemielniak, Masukume & Wilamowski, 2019). The significant coverage of other languages like French, Spanish, German, Chinese and Swedish in Medical Wikidata can be explained by their support by major biomedical multilingual databases: ICPC-2 (Rodgers, Sherwin, Lamberts, & Okkes, 2004) supports 24 languages[12], SNOMED-CT supports 7 languages, LOINC[13] supports 13 languages, BabelMeSH (Fontelo, Liu, Leon, Abrahamane, & Ackerman, 2007) supports 13 languages, and ICD-10 supports 42 languages. 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[14] (Terryn, Hoste, & Lefever, 2020) and by the involvement of research institutions and scientists speaking these languages, particularly German and Dutch, in adding biomedical information to Wikidata (Waagmeester, Schriml, & Su, 2019; Putnam, et al., 2017).

These factors are not the only ones behind the language coverage of medical Wikidata relations, as the distribution of medical Wikidata labels, particularly for diseases' class, seems to be linked in part to the number of speakers of each language among the editors of Wikidata (Kaffee & Simperl, 2018), to the overall number of Wikidata labels for each language (Kaffee, et al., 2017) and to the number of medical Wikipedia articles for each language (Heilman & West, 2015) as shown in Table 2.

Table 2: Languages ranked according to various variables, based on data 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 (Heilman & West, 2015)		Wikidata labels, 2017 (Kaffee, et al., 2017)		Population, 2019 (Eberhand, Simons, & Fennig, 2019)		Wikidata users, 2018 (Kaffee & Simperl, 2018)
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**

[12] 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.

[13] 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.

[14] 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.

10	Japanese	4303	Bengali	1.94%	**Western Punjabi**	82.5	Portuguese

These correlations can be put to a test by querying Wikidata to find out the current status of the editing of this knowledge graph and of Wikipedia, as shown in Table 3. Despite several differences like the higher visibility of Asian languages (particularly Arabic and Chinese) in that table, the query results largely match the literature-derived data reported in Table 2.

Table 3: Languages ranked according to various variables, based on 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 Desccriptor 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 correlate with Wikipedia pages, edits and pageviews related to COVID-19 pandemic as shown in Table 4. Consequently, encouraging the contribution by speakers of under-resourced and unrepresented languages to medical Wikipedia projects[15] and to Medical Wikidata is highly valuable to ameliorate and increase the language coverage of Wikidata in medical domains. However, this language coverage of the studied medical relations of Wikidata does not correlate with the number of speakers of each natural language as shown in Table 2. This finding aligns with previous research showing that the overall language distribution of Wikidata labels does not seem to correlate with the number of speakers of each language (Kaffee, et al., 2017).

Table 4: Languages ranked according to various COVID-19-related variables (as of August 13, 2020). The COVID Wikidata content query sorts languages by the number of labels of Wikidata items with a direct link to and/or from any of the core

[15] 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.

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 pages 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. The COVID-19 pandemic Wikipedia pageviews column represents daily average user traffic (averaged since January 1, 2020) to the article about the COVID-19 pandemic in the respective language. Style code: *Italic* for languages appearing in all four lists; **bold** for those appearing in only one.

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

4. 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 (Turki, et al., 2020a). As of September 1, 2020, 5302[16] out of 7877[17] Wikidata properties are used to state the external identifiers of the Wikidata items. These properties ensure the 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 (Vrandečić & Krötzsch, 2014; Färber, et al., 2018; Erxleben, et al., 2014). 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 (Vrandečić & Krötzsch, 2014; Ristoski, De Vries, & Paulheim, 2016) and for the integration of scholarly knowledge (Mietchen, et al., 2015).

In the circumstances of the COVID-19 outbreak, a SPARQL query[18] has been formulated to analyze the integration of external identifiers in Wikidata. This query

[16] For the updated count of the properties defining external identifiers, refer to https://w.wiki/ayn.
[17] For the updated count of all the properties, refer to https://w.wiki/ayo.
[18] https://w.wiki/auR

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 Wikidata within the linked open data cloud (Fig. 3) and consequently the usefulness of Wikidata to dynamically integrate various types of semantic data in the context of the disease outbreak (Debattista, et al., 2018).

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 (Nielsen, Mietchen, & Willighagen, 2017; Mietchen, 2020; Wyatt, et al., 2020). 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[19] 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. In fact, 51007 COVID-19-related records are available on PubMed[20], 117583 COVID-19 publications are available on Dimensions[21], 53001 are accessible on PubMed Central[22], 213000 records are available on Semantic Scholar[23], 3232 records are found at ClinicalTrials.gov[24], 2017 records are available on arXiv ID[25], and 53 records are reachable on NIOSHTIC-2[26] as of September 3, 2020. This lack of support is mainly explained by the method of development of scientific metadata on Wikidata that is 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 (Taraborelli, et al., 2017; Wyatt, et al., 2020).

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

[19] https://scholia.toolforge.org/
[20] https://pubmed.ncbi.nlm.nih.gov/?term=COVID-19
[21] https://tinyurl.com/y6kwrdth
[22] https://www.ncbi.nlm.nih.gov/pmc/?term=COVID-19
[23] https://www.semanticscholar.org/search?q=COVID-19&sort=relevance
[24] https://clinicaltrials.gov/ct2/results?cond=COVID-19&term=&cntry=&state=&city=&dist=
[25] https://arxiv.org/search/?query=COVID-19&searchtype=all&source=header
[26] https://www2a.cdc.gov/nioshtic-2/Buildqyr.asp?S1=COVID-19&Submit=Search

P2536	Sandbox-External identifier	6790
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 interestingly assigned them external identifiers in the main biomedical semantic databases such as MeSH, ICD-10[27], and UMLS[28] 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 (Rahimi, et al., 2020). 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 (Waagmeester, et al., 2020b). Diseases and symptoms are also assigned external identifiers linking them to several online encyclopedias (e.g. eMedicine, Encyclopedia Britannica, and MedlinePlus) and to non-medical databases such as scholarly repositories (e.g. JSTOR[29]) and bibliographic databases (e.g. Microsoft Academic[30]). This can be explained by the efforts of WikiProject WikiCite to align topic pages in research databases to Wikidata items so that the 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 (Nielsen, Mietchen, & Willighagen, 2017). The links between diseases and symptoms to online first-class encyclopedias is not restricted to the context of COVID-19 pandemic (Waagmeester, et al., 2020b) and is rather established to provide further specialized readings to users about every concerned Wikidata item (Klein & Kyrios, 2013) and to allow comparison of medical data quality between Wikipedia and other encyclopedias (Heilman & West, 2015).

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[31]	32	1
P3827	JSTOR topic ID	32	10
P6366	Microsoft Academic ID	29	11

[27] International Classification of Diseases, Tenth Revision (https://www.who.int/classifications/icd/en/)
[28] Unified Medical Language System (https://www.nlm.nih.gov/research/umls/index.html)
[29] https://www.jstor.org/
[30] https://academic.microsoft.com/
[31] International Classification of Diseases, Tenth Revision, Clinical Modification

P493	ICD-9[32]	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[33] Thesaurus ID	17	7
P1296	Gran Enciclopedia Catalana ID	10	7
P8408	KBpedia[34] ID	16	7

The matching between Wikidata items and online encyclopedias and non-medical resources is not only restricted to disease and symptoms but also to humans and sovereign states (e.g. VIAF, WorldCat, Library of Congress and GeoNLP) as shown in Table 7 and to films (e.g. IMDb), computer applications (e.g. Google Play) and disease outbreaks (e.g. Subreddit) as shown in Table 8. The alignment to various metadata databases like VIAF[35], WorldCat[36], Library of Congress and IMDb[37] 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 (Klein & Kyrios, 2013; Allison-Cassin & Scott, 2018). The alignment of Wikidata items about sovereign states and humans to corresponding topic pages and user pages in social networking services (Twitter) and question answering forums (Quora and Reddit) are mainly done to track the effect of the information provided by Wikimedia projects, particularly Wikipedia, on online communities (Vincent, Johnson, & Hecht, 2018) or to retrieve information about items in social media and consequently to support the topic modelling of the coverage of COVID-19 pandemic in social networks (Ciechanowski, Jemielniak, & Gloor, 2020). The sum of 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 (Mietchen, et al., 2015).

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

[32] International Classification of Diseases, Ninth Revision
[33] Biblioteca Nazionale Centrale di Firenze (Central National Library of Florence, Italy)
[34] https://kbpedia.org/
[35] Virtual International Authority File (http://viaf.org/)
[36] https://www.worldcat.org/
[37] Internet Movie Database (https://www.imdb.com/)

P244	Library of Congress authority ID	125	458
P213	ISNI[38]	100	443
P646	Freebase ID	124	379
P2002	Twitter username	16	353
P227	GND[39] ID	125	308
P345	IMDb ID		274
P268	Bibliothèque nationale de France ID	177	269
P269	IdRef[40] ID	84	265
P998	DMOZ[41] 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[42] 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[43]), for biodiversity (e.g. IRMNG[44]), for genomics (e.g. Entrez Gene) and for proteomics (e.g. PDB[45]) 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 clearly explicated by the better interest of the users of social media to the health policies and epidemiology of COVID-19 rather than the therapeutic options and molecular aspects related to the disease (Ordun, Purushotham, & Raff, 2020). The most important interest in matching these concepts in Wikidata to graph databases (e.g. Massbank, PDB, and KEGG[46]) and semi-structured databases (e.g. Guide to Pharmacology) for bioinformatics rather than online encyclopedias is due to the better availability of genomic and proteomic information in these specialized semantic resources (Huss III, et al., 2008; Waagmeester, et al., 2020b). The alignment of taxon items in Wikidata to biodiversity knowledge graphs (e.g. NCBI[47] 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 (Page, 2016). The sum of these biomedical alignments is developed using human edits and computer tools thanks to large initiatives to develop open

[38] https://isni.org/
[39] Gemeinsame Normdatei (German National Library, Germany), https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html
[40] Identifiants et Référentiels pour l'enseignement supérieur et la recherche (Identifiers and credentials for higher education and research in France)
[41] Directory Mozilla (https://dmoz-odp.org/)
[42] https://www.loc.gov/marc/
[43] https://massbank.eu/MassBank/
[44] Interim Register of Marine and Nonmarine Genera (https://www.irmng.org/)
[45] Protein Data Bank (https://www.rcsb.org/)
[46] Kyoto Encyclopedia of Genes and Genomes (https://www.genome.jp/kegg/)
[47] National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/)

ontological databases for curating advanced molecular biology data such as WikiGenomes (Putman, et al., 2017) and Gene Wiki (Burgstaller-Muehlbacher, et al., 2016) and is enhanced in the context of the current pandemic through the contributions of WikiProject COVID-19 (Waagmeester, et al., 2020a).

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[48]	31
protein [Q8054]	P638	PDB structure ID	31
film [Q11424]	P345	IMDb ID	25
film [Q11424]	P2603	Kinopoisk film ID	23
film [Q11424]	P7177	Cinestaan film ID	22
disease outbreak [Q3241045]	P3984	subreddit	22
protein [Q8054]	P637	RefSeq[49] protein ID	18
committee group motion [Q97695005]	P8433	Swedish Riksdag document ID	18
film [Q11424]	P2529	ČSFD[50] film ID	17
drug [Q11173]	P267	ATC[51] 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[52] 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[53]	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

[48] Spectral Hash Identifier (https://splash.fiehnlab.ucdavis.edu/)
[49] NCBI Reference Sequence Database (https://www.ncbi.nlm.nih.gov/refseq/)
[50] Česko-Slovenská filmová databáze (Czech-Slovak Film Database, https://www.csfd.cz/)
[51] Anatomical Therapeutic Chemical (ATC) Classification System (https://www.whocc.no/atc_ddd_index/)
[52] https://www.cas.org/support/documentation/chemical-substances
[53] Unique Ingredient Identifier (https://fdasis.nlm.nih.gov/srs/)

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

5. 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[54] can illustrate different aspects of the COVID-19 disease, its causative virus, and the resulting pandemic (extended list, supplementary table S2). Sample SPARQL queries are available at supplementary table S1. This section will present examples across different aspects of COVID-19, adapted from five main sources to which we have contributed substantially[55,56,57,58,59].

Biological and clinical aspects

A simple demonstration of Wikidata's encoding of SARS-CoV-2's basic biology is in its genetics (Fig. 5) and resulting symptoms (Fig. 6). 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 (Putnam, et al., 2017; Manske, et al., 2019), 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) (Waagmeester, et al., 2020b). 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).

---

[54] Technical documentation about SPARQL can be found at https://en.wikibooks.org/wiki/SPARQL.
[55] 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
[56] SARS-CoV-2-Queries: extracts from the book "Wikidata Queries around the SARS-CoV-2 virus and pandemic" (Addshore, Mietchen & Willighagen, 2020); https://egonw.github.io/SARS-CoV-2-Queries/
[57] SPEED queries: extracts from the Wikidata-based epidemiological surveillance dashboard for COVID-19 pandemic in Tunisia. It was partially built upon COVID-19 Wikidata dashboard; https://sites.google.com/view/covid19-dashboard
[58]Scholia queries: queries underlying COVID-19-related visualizations from the Wikidata-based scholarly profiling tool Scholia (Nielsen, Mietchen, & Willighagen, 2017); https://scholia.toolforge.org/
[59]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

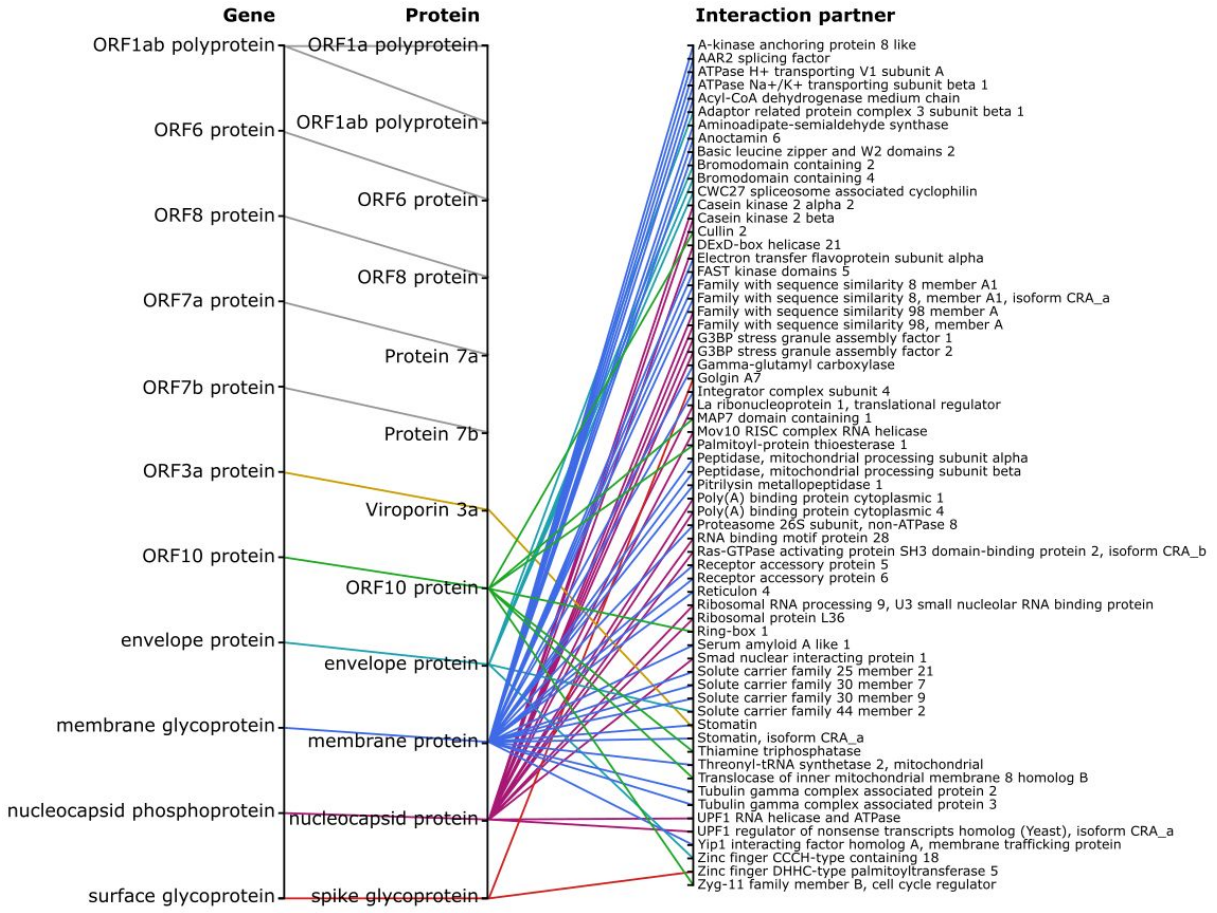
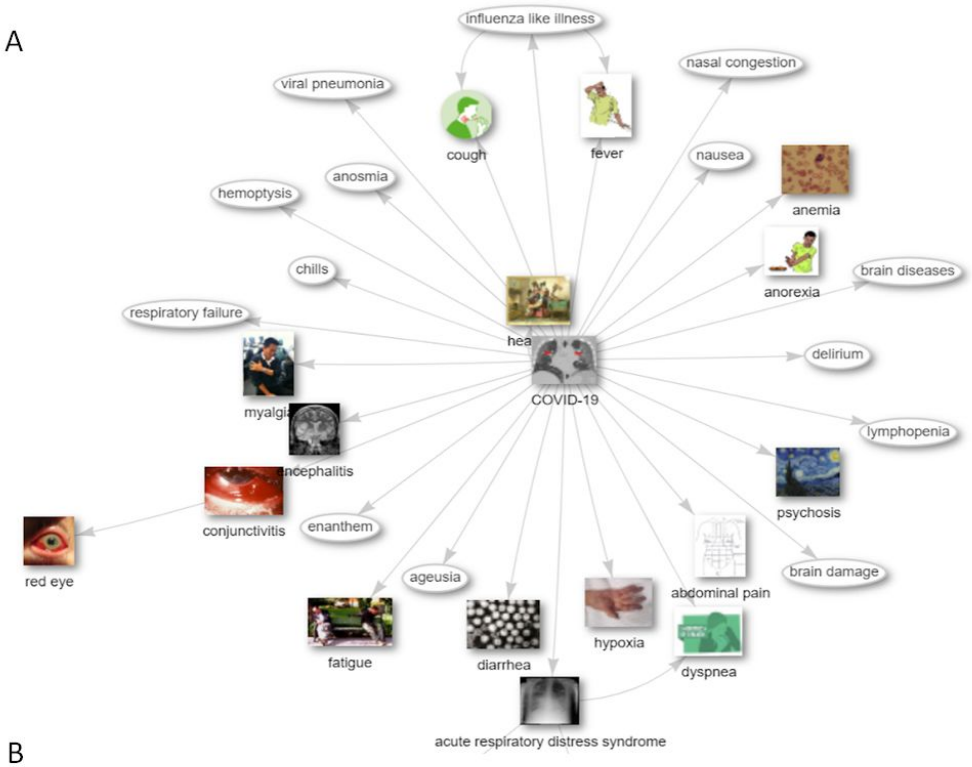


Figure 5. SARS-CoV-2 interactions with the human proteome (available at: https://w.wiki/c3D, license: CC BY 4.0). Proteins encoded by SARS-CoV-2 genes (note that some genes encode multiple proteins) and the currently known human protein interaction partners (live data: https://w.wiki/beR).

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

Figure 6: Symptoms of COVID-19 and similar conditions (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)

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. 7A), whereas others common visualisations can be calculated from scratch from granular data such as the individual *date of birth* [P569] and *of death* [P570] of notable individuals deceased from COVID-19 (Figure 7B). Although this reflects the age distribution of COVID mortality, it is also influenced by the demographics of persons sufficiently notable to have Wikidata items. In some cases summary data is also time-resolved, allowing inquiry of its

change over time over time (supplementary figure S2), capturing features not depicted in several statistical predictions of the epidemiological evolution of COVID-19 outbreaks (Chaari & Golubnitschaja, 2020) and clearly seen in other data sources, such that mortality peaks at the beginning of a disease outbreak (Zhang, et al., 2020). Wikidata's granularity 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, a result likely difficult to achieve with other datasets (supplementary figure S3), and mirroring Wikipedia's well-known rapid response to updating information on deaths (Keegan & Brubaker, 2015; Keegan & Tan, 2020).

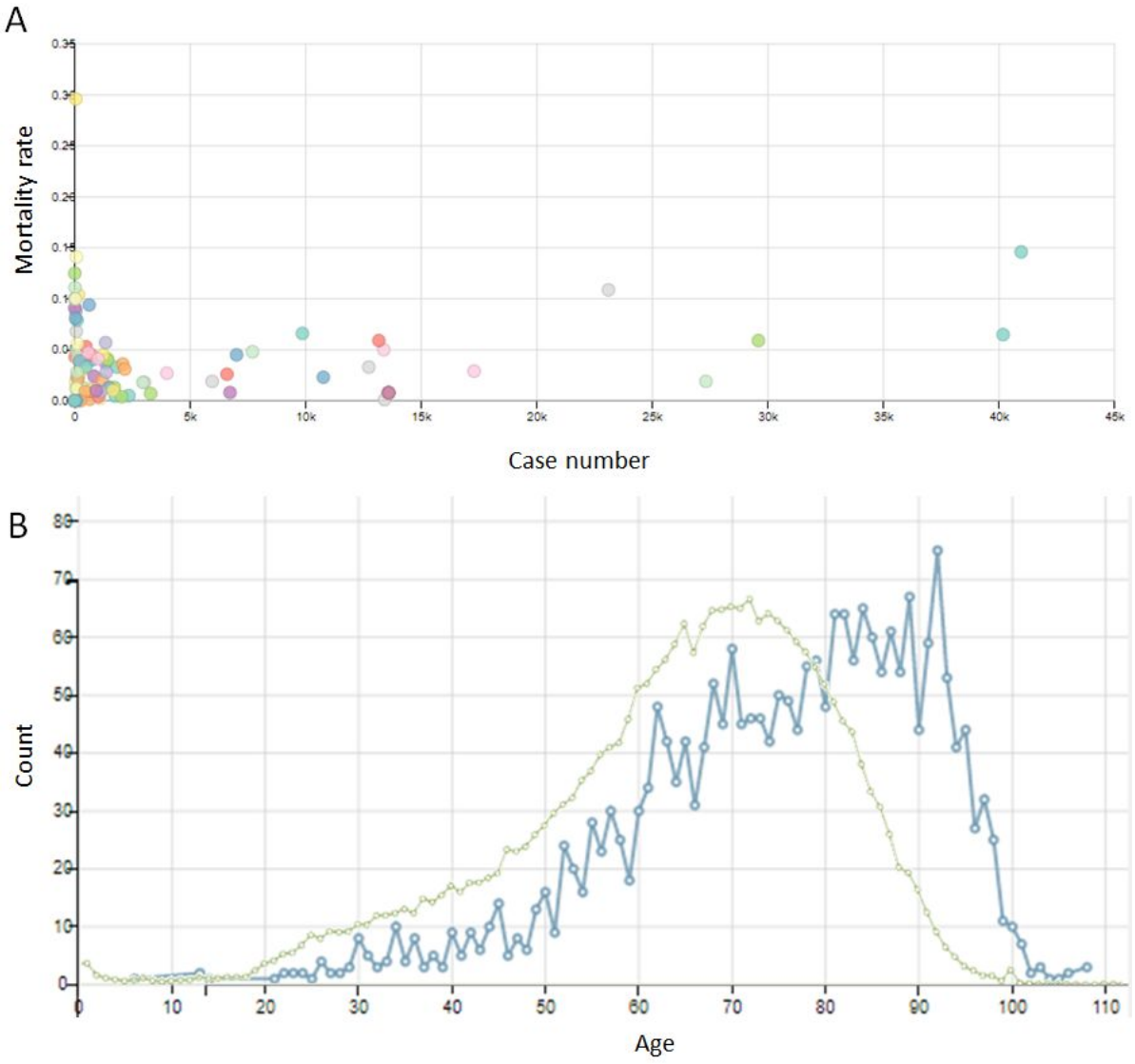


Figure 7. Summary epidemiological data on the COVID-19 pandemic (available at: https://w.wiki/byW, license: CC BY 4.0). A) Correlation between the current number of cases and mortality rates in every country, calculated from numeric summary data for each region. Countries coloured randomly (live data: https://w.wiki/bf$). B) Age distribution of notable persons who have died of COVID-19 (blue), compared to the death age distribution for people 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).

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. 8A). This is a useful and rapid first-approximation of topics covered by those publications, extracted as plain text, and can be expanded upon by querying the main topics to concepts in Wikidata using the property *main subject* (P921). This property acts analogously to Medical Subject Headings (MeSH) descriptors (Turki, Hadj Taieb, & Ben Aouicha, 2018), though far more extensive. Those statements allow broader querying of the literature as a network via co-occurrence of topics as the main subject of articles (Fig. 8B)[60]. This enables rapid traversal and faceting of the literature on topics in addition to the traditional links made by tracing citations (Hu, Rousseau, & Chen, 2011), 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) (Boccone & Rivelli, 2019).

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 (Majumder & Mandl, 2020). Further, Wikidata's rich biographical and institutional data makes extracting information on authors or others easy (Fig. S8), and eventually other contributors (Nunn, et al., 2019).

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

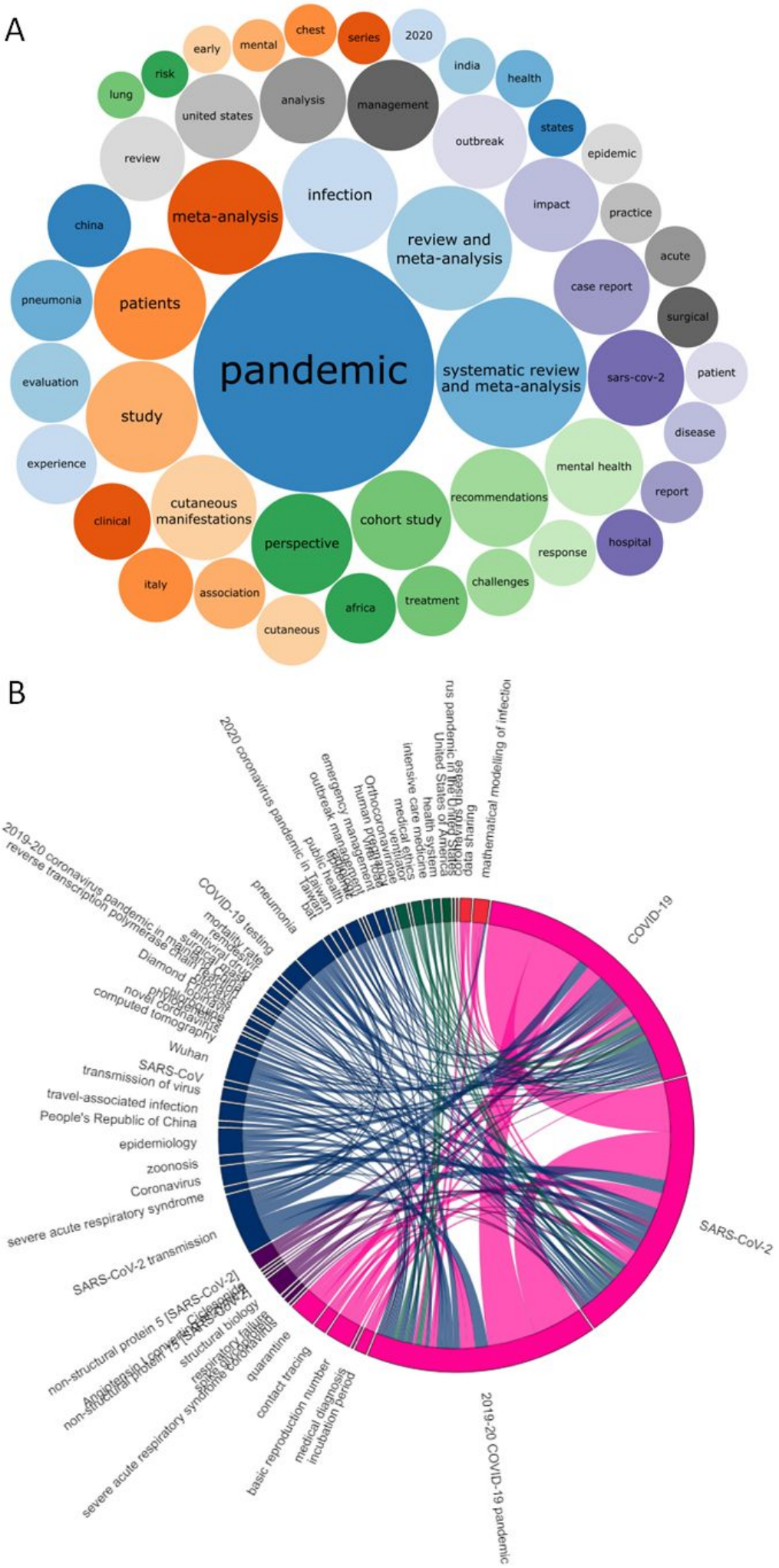

A
pandemic
meta-analysis
infection
review and meta-analysis
systematic review and meta-analysis
patients
study
cutaneous manifestations
perspective
cohort study
review
china
pneumonia
evaluation
experience
clinical
italy
association
cutaneous
africa
treatment
challenges
response
recommendations
mental health
hospital
report
disease
patient
sars-cov-2
surgical
acute
case report
practice
impact
epidemic
outbreak
states
health
india
2020
series
chest
mental
early
risk
lung
united states
analysis
management
B
COVID-19
SARS-CoV-2
2019-20 COVID-19 pandemic
pneumonia
2020 coronavirus pandemic in Taiwan
bat
public health
emergency management
outbreak
human coronavirus
Orthocoronavirinae
ventilator
medical ethics
intensive care medicine
health system
United States of America
mathematical modelling of infectio
COVID-19 testing
mortality rate
remdesivir
antiviral drug
surgical mask
Diamond Princess
chloroquine
novel coronavirus
computed tomography
Wuhan
SARS-CoV
transmission of virus
travel-associated infection
People's Republic of China
epidemiology
zoonosis
Coronavirus
severe acute respiratory syndrome
SARS-CoV-2 transmission
non-structural protein 5 [SARS-CoV-2]
structural biology
respiratory failure
quarantine
contact tracing
basic reproduction number
medical diagnosis
incubation period

Figure 8. COVID-19 publication topics (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).

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. 9), 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. Standardised methods to audit and validate Wikidata's content on various topics are still under investigation and development (Turki, et al., 2020a).

A

organization	organizationLabel	bankruptcyDate	countryLabel	inception	industries	parents	subsiduaries
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		

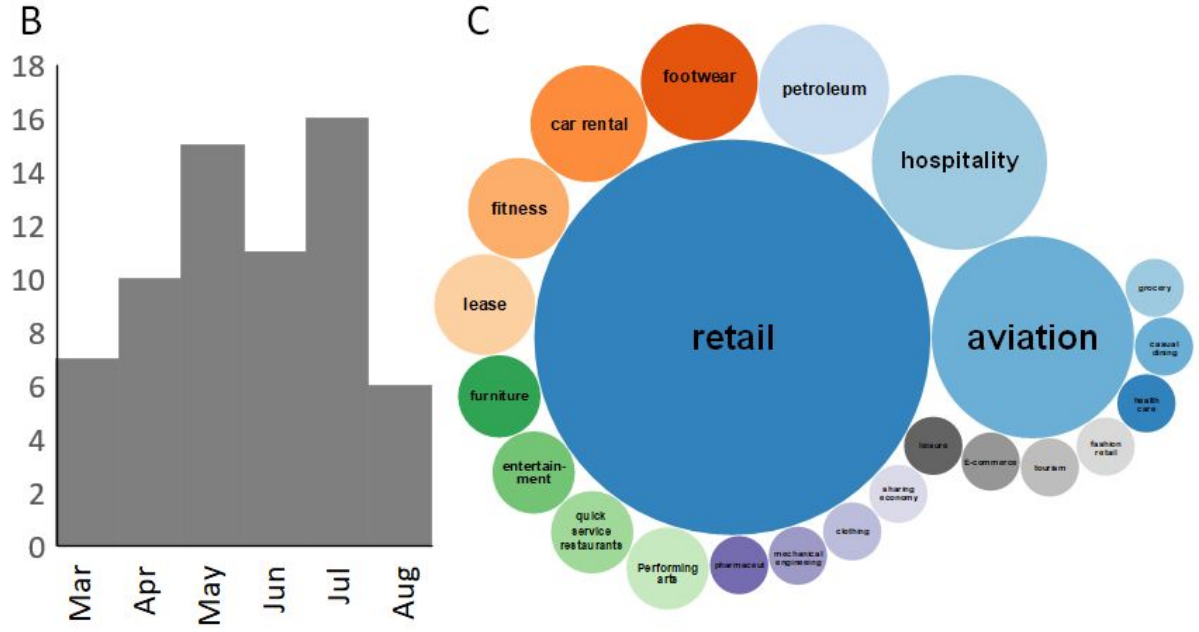


Figure 9. Bankrupt publicly listed businesses due to the COVID-19 pandemic (available at: https://w.wiki/byY, license: CC BY 4.0). A) Tabular output of SPARQL query B) Bankruptcies per month C) ratios of different industries associated with bankrupt companies. (live data:https://w.wiki/cG6).

6. Discussion

Many knowledge graphs have been recently developed to represent various types of COVID-19 information, including government responses (Desvars-Larrive, et al., 2020), epidemiology (Xu, Kraemer, & Data Curation Group, 2020), clinical data (Ostaszewski, et al., 2020), scholarly outputs and outcomes (Jaradeh, et al., 2019), economic impacts (Bellomarini, et al., 2020), physiopathology (Domingo-Fernández, et al., 2020), social networking (Dimitrov, et al., 2020) among other features related to the COVID-19 pandemic. The semantic databases are either built using a combination of human efforts and crowdsourcing techniques (Desvars-Larrive, et al., 2020) or through the automatic extraction

of information using natural language processing techniques from scholarly publications about the outbreak, particularly the Covid-19 Open Research Dataset (Wang, et al., 2020). Despite the importance of these provided resources, they tend to cover a narrow aspect of the disease. That is why initiatives have been launched by closed groups of data curators to create and sustain platforms to integrate COVID-19 knowledge such as CIDO (He, et al., 2020) and COVID-19 data hub (Guidotti & Ardia, 2020). These centralized semantic databases allowed the combination of many divergent factors of COVID-19 (e.g., clinical data with genomics, and epidemiological data with social interactions) to achieve more complex scientific interpretations about SARS-CoV-2 virus and the pandemic and consequently to support more advanced decision making related to the outbreak (He, et al., 2020; Guidotti & Ardia, 2020). However, due to the sharply growing nature of the scholarly literature about coronaviruses (Kagan, Moran-Gilad, & Fire, 2020), the COVID-19 knowledge has become difficult to follow and manage by a restricted community of scientists.

Whereas most knowledge graphs tend to be specialized and closed, Wikidata deliberately takes a multidisciplinary, multilingual position 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, uniquely suited to topics such as the COVID-19 pandemic (Waagmeester, et al., 2020a; Waagmeester, et al., 2020b). 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 (Lehmann, et al., 2015). Wikidata is enriched and adjusted as well by a community of over 23000 active human users on a daily basis[61] and is released under the CC0 license allowing the free and unconditioned 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 (Turki, et al., 2020a). By being highly multilingual, its human-readability extends beyond just English to support international contribution (Turki, et al., 2019; Turki, et al., 2020a). 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 (Turki, et al., 2019; Turki, et al., 2020a).

These factors have facilitated Wikidata's rapid growth since its creation in 2012 - likely surpassing 90 million items in the coming months - into a richly interconnected and interdisciplinary network (Turki, et al., 2019; Turki, et al., 2020a). In the context of COVID-19 outbreak, Wikidata has proven its efficiency to represent multiple facets of the pandemic ranging from biomedical information to social impacts, surpassing other integrated semantic graphs such as CIDO (He, et al., 2020), COVID-19 data hub (Guidotti & Ardia, 2020), COVID-19 Living Data[62] (Juul, et al., 2020) and Knowledge4COVID-19[63] (Iglesias, et al., 2020) that only combine two to three distinct features of the pandemic 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

[61] https://www.wikidata.org/wiki/Special:Statistics
[62] https://covid-nma.com/
[63] https://devpost.com/software/covid-19-kg

revealed by “database alignment” section. Moreover, SPARQL as a semantic query language has been proved as useful to enable deeper insights into the different facets of the multidisciplinary COVID-19 information in Wikidata, characterized as big data by its volume, variety, velocity and veracity (Hitzler & Janowicz, 2013; Sebei, Hadj Taieb, & Ben Aouicha, 2018). This confirms previous findings about the importance of querying COVID-19 semantic resources such as CIDO (He, et al., 2020) to compare clinical information with other types of COVID-19 information and consequently to generate deeper characteristics of the disease (Dadzie & Rowe, 2011). The advantage of applying SPARQL to extract and visualize COVID-19 information from a generalized knowledge graphs such as WIkidata when compared to knowledge graphs developed for the pandemic like CIDO (He, et al., 2020) is that open knowledge graphs also include 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, The application of CC0 license 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 (Hagedorn, et al., 2011; Penev, et al., 2017). This situation seems analogous to the group O red blood cells status as universal donor but restricted recipient (Lublin, 2000).

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 misleading results if not interpreted with caution. For example, the COVID-19 outbreaks on cruise[64] and naval[65] ships are more covered in Wikipedia than in Wikidata (as well as in many other online resources). Another example can be the distorted representation of scholarly citations in Wikidata, since no scalable workflows currently exist for their systematic curation. 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 solved by applying algorithms that assess data completeness of an item included in a given class within open knowledge graphs by comparing it with other items of the same class (Darari, et al., 2016; Balaraman, Razniewski, & Nutt, 2018) or through enhancing the use of knowledge graph learning techniques to enrich, sustain and update open knowledge graphs from textual databases like scholarly publications (Zhang, et al., 2018) and electronic health records (Rotmensch, et al., 2017). Moreover, collaborative editing can cause several inaccuracies in the declaration of statements in open knowledge graphs disregarding the metadata standards of the

[64] https://en.wikipedia.org/wiki/COVID-19_pandemic_on_cruise_ships
[65] https://en.wikipedia.org/wiki/COVID-19_pandemic_on_naval_ships

knowledge bases (Schriml, et al., 2020). 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[66] and can consequently cause matters about the trustworthiness of the reuse of data (Schriml, et al., 2020). 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 (Färber, et al., 2018). This is not only due to the involvement of more contributors in Wikidata than in other open knowledge graphs (Färber, et al., 2018) or on the reliance of this knowledge base on secondary databases that are timely curated and enriched, mainly the ones from the linked open data cloud (Turki, et al., 2019) but also on the use of validation schemas in ShEx[67], of logical constraints implemented in SPARQL and of bot edits to verify the structure and accuracy of COVID-19 information (Turki, et al., 2020a; Farda-Sarbas, et al, 2019). The data validation infrastructure of Wikidata seems to be in accordance with the latest updates in knowledge graph evaluation and refinement techniques (Zaveri, et al., 2016; Paulheim, 2017) and explains in part the reasons behind the robustness of the data model of COVID-19 information in this open knowledge graph.

7. Conclusion

In this research paper, we demonstrate the ability of open and collaborative knowledge graphs such as Wikidata to represent and integrate all the multidisciplinary aspects of the COVID-19 information and to generate summary visualizations about the infectious disease using SPARQL. 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. An important partner for Wikidata FAIRness is the SPARQL query service. The SPARQL endpoint, combined with community-contributed data visualization tools, provides a human-friendly interface. 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.

Even though COVID-19 knowledge is abundant on Wikidata, there is always room for future improvement. As a collaborative endeavour, Wikidata (and the WikiProject COVID-19 specially) is likely to become richer with 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. 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

[66] As of September 13, 2020, there are only 63 Wikidata administrators as shown at https://www.wikidata.org/wiki/Special:Statistics.

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

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.

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 (Turki, et al., 2020b). 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[68] 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 (Aturban, et al., 2018).

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. 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

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

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

Aturban, M., Kelly, M., Alam, S., Berlin, J. A., Nelson, M. L., & Weigle, M. C. (2018, May). ArchiveNow: Simplified, Extensible, Multi-Archive Preservation. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (pp. 321-322). doi:10.1145/3197026.3203880.

[68] 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.

Balaraman, V., Razniewski, S., & Nutt, W. (2018, April). Recoin: relative completeness in Wikidata. In *Companion Proceedings of the The Web Conference 2018* (pp. 1787-1792). doi:10.1145/3184558.3191641.

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

Boccone, A., & Rivelli, R. (2019). The bibliographic metadata in Wikidata: Wikicite and the «Bibliothecae.it» case study. *Bibliothecae.it*, *8*(1), 227-248. doi:10.6092/issn.2283-9364/9503.

Burgstaller-Muehlbacher, S., Waagmeester, A., Mitraka, E., Turner, J., Putman, T., Leong, J., . . . Su, A. I. (2016). Wikidata as a semantic framework for the Gene Wiki initiative. *Database, 2016*, baw015. doi:10.1093/database/baw015.

Chaari, L., & Golubnitschaja, O. (2020). 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), 133-138. doi:10.1007/s13167-020-00207-0.

Ciechanowski, L., Jemielniak, D., & Gloor, P. A. (2020). TUTORIAL: AI research without coding: The art of fighting without fighting: Data science for qualitative researchers. *Journal of Business Research*, *117*, 322-330. doi:10.1016/j.jbusres.2020.06.012.

Dadzie, A. S., & Rowe, M. (2011). Approaches to visualising linked data: A survey. *Semantic Web*, *2*(2), 89-124. doi:10.3233/SW-2011-0037.

Darari, F., Razniewski, S., Prasojo, R. E., & Nutt, W. (2016). Enabling fine-grained RDF data completeness assessment. In *International Conference on Web Engineering* (pp. 170-187). Springer, Cham. doi:10.1007/978-3-319-38791-8_10.

De Melo, G., & Weikum, G. (2010). Towards universal multilingual knowledge bases. In *Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the 5th Global WordNet Conference (GWC 2010)* (pp. 149-156). Narosa Publishing, New Delhi, India.

Debattista, J., Lange, C., Auer, S., & Cortis, D. (2018). Evaluating the quality of the LOD cloud: An empirical investigation. *Semantic Web*, *9*(6), 859-901. doi:10.3233/SW-180306.

Desvars-Larrive, A., Dervic, E., Haug, N., Niederkrotenthaler, T., Chen, J., Di Natale, A., et al. (2020). A structured open dataset of government interventions in response to COVID-19. *Scientific data, 7*(1), 285. doi:10.1038/s41597-020-00609-9.

Dimitrov, D., Baran, E., Fafalios, P., Yu, R., Zhu, X., Zloch, M., & Dietze, S. (2020). TweetsCOV19--A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. *arXiv preprint arXiv:2006.14492*.

Domingo-Fernández, D., Baksi, S., Schultz, B., Gadiya, Y., Karki, R., Raschka, T., et al. (2020, April 15). COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *BioRxiv*. doi:10.1101/2020.04.14.040667.

Eberhand, D. M., Simons, G. F., & Fennig, C. D. (2019). *Ethnologue: Languages of the World.* Dallas, Texas: SIL International.

Ehrlinger, L., & Wöß, W. (2016). Towards a Definition of Knowledge Graphs. *CEUR Workshop Proceedings, 1695*, 1-4.

Ekins, S., Mietchen, D., Coffee, M., Stratton, T. P., Freundlich, J. S., Freitas-Junior, L., ... & Andrade, C. (2016). Open drug discovery for the Zika virus [version 1; peer review: 3 approved]. *F1000Research, 2016*, 5:150. doi: 10.12688/f1000research.8013.1.

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., & Vrandečić, D. (2014, October). Introducing Wikidata to the linked data web. In *International semantic web conference* (pp. 50-65). Springer, Cham. doi:10.1007/978-3-319-11964-9_4.

Färber, M., Bartscherer, F., Menne, C., & Rettinger, A. (2018). Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web, 9*(1), 77–129. doi:10.3233/SW-170275.

Farda-Sarbas, M., Zhu, H., Nest, M. F., & Müller-Birn, C. (2019). Approving automation: analyzing requests for permissions of bots in Wikidata. In *Proceedings of the 15th International Symposium on Open Collaboration* (pp. 1-10). doi:10.1145/3306446.3340833.

Fontelo, P., Liu, F., Leon, S., Abrahamane, A., & Ackerman, M. (2007). PICO Linguist and BabelMeSH: development and partial evaluation of evidence-based multilanguage search tools for MEDLINE/PubMed. *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics* (p. 817). IOS Press.

Freitas, F., Schulz, S., & Moraes, E. (2009). Survey of current terminologies and ontologies in biology and medicine. *Reciis, 3*(1), 7-18. doi:10.3395/reciis.v3i1.239en.

Guidotti, E., & Ardia, D. (2020). COVID-19 data hub. *Journal of Open Source Software*, *5*(51), 2376. doi:10.21105/joss.02376.

Hagedorn, G., Mietchen, D., Morris, R. A., Agosti, D., Penev, L., Berendsohn, W. G., & Hobern, D. (2011). Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. *ZooKeys*, (150), 127. doi:10.3897/zookeys.150.2189.

He, Y., Yu, H., Ong, E., Wang, Y., Liu, Y., Huffman, A., et al. (2020). CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. *Scientific data, 7*(1), 181. doi:10.1038/s41597-020-0523-6.

Henriksson, A., Skeppstedt, M., Kvist, M., Duneld, M., & Conway, M. (2013, August). 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* (pp. 36-44).

Heilman, J. M., & West, A. G. (2015). Wikipedia and medicine: quantifying readership, editors, and the significance of natural language. *Journal of Medical Internet Research, 17*(3), e62. doi:10.2196/jmir.4069.

Hitzler, P., & Janowicz, K. (2013). Linked Data, Big Data, and the 4th Paradigm. *Semantic Web*, *4*(3), 233-235. doi:10.3233/SW-130117.

Hu, X., Rousseau, R., & Chen, J. (2011). On the definition of forward and backward citation generations. *Journal of Informetrics*, *5*(1), 27-36. doi:10.1016/j.joi.2010.07.004.

Huss III, J. W., Orozco, C., Goodale, J., Wu, C., Batalov, S., Vickers, T. J., ... & Su, A. I. (2008). A gene wiki for community annotation of gene function. *PLoS Biol*, *6*(7), e175.

Iglesias, E., Jozashoori, S., Chaves-Fraga, D., Collarana, D., & Vidal, M. E. (2020). SDM-RDFizer: An RML Interpreter for the Efficient Creation of RDF Knowledge Graphs. *arXiv preprint arXiv:2008.07176*.

Jaradeh, M. Y., Oelen, A., Farfar, K. E., Prinz, M., D'Souza, J., Kismihók, G., ... & Auer, S. (2019). Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture* (pp. 243-246). doi:10.1145/3360901.3364435.

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

Jemielniak, D., Masukume, G., & Wilamowski, M. (2019). The most influential medical journals according to Wikipedia: quantitative analysis. *Journal of medical Internet research*, *21*(1), e11429. doi:10.2196/11429.

Jetté, N., Quan, H., Hemmelgarn, B., Drosler, S., Maass, C., Oec, D.-G., . . . Ghali, W. A. (2010). The development, evolution, and modifications of ICD-10: challenges to the international comparability of morbidity data. *Medical Care, 48*(12), 1105-1110. doi:10.1097/MLR.0b013e3181ef9d3e.

Juul, S., Nielsen, N., Bentzer, P., Veroniki, A. A., Thabane, L., Linder, A., ... & Jakobsen, J. C. (2020). 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*, 108. doi:10.1186/s13643-020-01371-0.

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

Kaffee, L.-A., & Simperl, E. (2018). Analysis of Editors' Languages in Wikidata. *Proceedings of the 14th International Symposium on Open Collaboration* (p. 21). ACM. doi:10.1145/3233391.3233965

Kagan, D., Moran-Gilad, J., & Fire, M. (2020). Scientometric trends for coronaviruses and other emerging viral infections. *GigaScience, 9*(8), giaa085. doi:10.1093/gigascience/giaa085.

Keegan, B. C., & Brubaker, J. R. (2015). '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* (pp. 533-546). doi:10.1145/2675133.2675238.

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

Kirk, A. (2016). *Data visualisation: A handbook for data driven design*. Sage. ISBN:9781473966314

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

Konieczny, P. (2010). Adhocratic governance in the Internet age: A case of Wikipedia. *Journal of Information Technology & Politics*, *7*(4), 263-283. doi:10.1080/19331681.2010.489408.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... & Bizer, C. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, *6*(2), 167-195. doi:10.3233/SW-140134.

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

Liu, Y., Chan, W. K. B., Wang, Z., Hur, J., Xie, J., Yu, H., & He, Y. (2020). Ontological and bioinformatic analysis of anti-coronavirus drugs and their Implication for drug repurposing against COVID-19. *Preprints, 2020*, 2020030413. doi:10.20944/preprints202003.0413.v1.

Lublin, D. M. (2000). Universal RBCs. *Transfusion*, *40*(11), 1285-1289. doi:10.1046/j.1537-2995.2000.40111285.x.

Majumder, M. S., & Mandl, K. D. (2020). Early in the epidemic: impact of preprints on global discourse about COVID-19 transmissibility. *The Lancet Global Health*, *8*(5), e627-e630. doi:10.1016/S2214-109X(20)30113-3.

Malyshev, S., Krötzsch, M., González, L., Gonsior, J., & Bielefeldt, A. (2018). Getting the most out of Wikidata: Semantic technology usage in wikipedia's knowledge graph. *International Semantic Web Conference* (pp. 376-394). Springer, Cham. doi:10.1007/978-3-030-00668-6_23.

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

Mietchen, D., Hagedorn, G., Willighagen, E., Rico, M., Gómez-Pérez, A., Aibar, E., Rafes, K,, Germain, C., Dunning, A., Pintscher, L., & Kinzler, D. (2015). Enabling open science: Wikidata for research (Wiki4R). *Research Ideas and Outcomes*, *1*, e7573. doi:10.3897/rio.1.e7573.

Mietchen, D. (2020). State of WikiCite in 2020. *Workshop On Open Citations And Open Scholarly Metadata 2020*. doi:10.5281/zenodo.4019954.

Nielsen, F. Å., Mietchen, D., & Willighagen, E. (2017). Scholia, scientometrics and Wikidata. In *European Semantic Web Conference* (pp. 237-259). Springer, Cham. doi:10.1007/978-3-319-70407-4_36.

Nunn, J., Shafee, T., Chang, S., Stephens, R., Elliott, J., Oliver, S., ... & Orr, N. (2019). Standardised Data on Initiatives-STARDIT: Alpha Version. *OSF Preprints*. doi:10.31219/osf.io/5q47h.

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

Ostaszewski, M., Mazein, A., Gillespie, M. E., Kuperstein, I., Niarakis, A., Hermjakob, H., ... & Schreiber, F. (2020). COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Scientific data*, *7*(1), 1-4. doi:10.1038/s41597-020-0477-8.

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

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web, 8*(3), 489-508. doi:10.3233/SW-160218.

Penev, L., Mietchen, D., Chavan, V., Hagedorn, G., Smith, V., Shotton, D., ... & Groom, Q. (2017). Strategies and guidelines for scholarly publishing of biodiversity data. *Research Ideas and Outcomes*, *3*, e12431. doi:10.3897/rio.3.e12431.

Putnam, T. E., Lelong, S., Burgstaller-Muehlbacher, S., Waagmeester, A., Diesh, C., Dunn, N., . . . Good, B. M. (2017). WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata. *Database, 2017*, bax025. doi:10.1093/database/bax025.

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

Ristoski, P., De Vries, G. K. D., & Paulheim, H. (2016, October). A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web. In *International Semantic Web Conference* (pp. 186-194). Springer, Cham. doi:10.1007/978-3-319-46547-0_20.

Rodgers, R. C., Sherwin, Z., Lamberts, H., & Okkes, I. M. (2004). ICPC Multilingual Collaboratory: a Web-and Unicode-based system for distributed editing/translating/viewing of the multilingual International Classification of Primary Care. *Medinfo 2004* (pp. 425-429). IOS Press. doi:10.3233/978-1-60750-949-3-425.

Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). Learning a health knowledge graph from electronic medical records. *Scientific reports*, *7*(1), 1-11. doi:10.1038/s41598-017-05778-z.

Schriml, L. M., Chuvochina, M., Davies, N., Eloe-Fadrosh, E. A., Finn, R. D., Hugenholtz, P., ... & Mizrachi, I. K. (2020). COVID-19 pandemic reveals the peril of ignoring metadata standards. *Scientific data*, *7*(1), 188. doi:10.1038/s41597-020-0524-5.

Sebei, H., Hadj Taieb, M. A., & Ben Aouicha, M. (2018). Review of social media analytics process and big data pipeline. *Social Network Analysis and Mining*, *8*(1), 30. doi:10.1007/s13278-018-0507-0.

Shafee, T., Masukume, G., Kipersztok, L., Das, D., Häggström, M., & Heilman, J. (2017). Evolution of Wikipedia's medical content: past, present and future. *J Epidemiol Community Health*, *71*(11), 1122-1129. doi:10.1136/jech-2016-208601.

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

Terryn, A. R., Hoste, V., & Lefever, E. (2019). In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation, 54*, 385-418. doi:10.1007/s10579-019-09453-9.

Turki, H., Hadj Taieb, M. A., & Ben Aouicha, M. (2018). MeSH qualifiers, publication types and relation occurrence frequency are also useful for a better sentence-level extraction of biomedical relations. *Journal of biomedical informatics, 83*, 217-218. doi:10.1016/j.jbi.2018.05.011.

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

Turki, H., Jemielniak, D., Hadj Taieb, M. A., Labra Gayo, J. E., Ben Aouicha, M., Banat, M., Shafee, T., Prud'Hommeaux, E., Lubiana, T., Das, D., & Mietchen, D. (2020a). Using logical constraints to

validate information in collaborative knowledge graphs: a study of COVID-19 on Wikidata. *Zenodo*. doi:10.5281/zenodo.4008358.

Turki, H., Hadj Taieb, M. A., Shafee, T., Lubiana, T., Jemielniak, D., Ben Aouicha, M., Labra Gayo, J. E., Banat, M., Das, D., & Mietchen, D. (2020b). csisc/WikidataCOVID19SPARQL: Data about Wikidata coverage of COVID-19. *Zenodo*. doi:10.5281/zenodo.4022591.

Vincent, N., Johnson, I., & Hecht, B. (2018). 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* (pp. 1-13). doi:10.1145/3173574.3174140.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78-85. doi:10.1145/2629489.

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

Waagmeester, A., Willighagen, E. L., Su, A. I., Kutmon, M., Labra Gayo, J. E., Fernández-Álvarez, D., et al. (2020a). A protocol for adding knowledge to Wikidata, a case report. *BioRxiv*. doi:10.1101/2020.04.05.026336.

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

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

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

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

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

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

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

Supplementary Data

Representing COVID-19 information in collaborative knowledge graphs: a study of Wikidata

Houcemeddine Turki, Mohamed Ali Hadj Taieb, Thomas Shafee, Tiago Lubiana, Dariusz Jemielniak, Mohamed Ben Aouicha, Jose Emilio Labra Gayo, Mus'ab Banat, Diptanshu Das, and Daniel Mietchen*, on behalf of WikiProject COVID-19

***Corresponding author:** Daniel Mietchen
School of Data Science, University of Virginia, Charlottesville, Virginia, United States of America
dm7gn@virginia.edu

Supplementary tables

Table S1. List of sample queries on COVID-19 overview, international situation, international daily epidemiological evolution, Tunisian daily epidemiological evolution,Tunisian governorate-level situation, Tunisian correlations, and Worldwide correlations.

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

Table S2. 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)

Supplementary figures

This section of the supplementary data includes additional array of visualisations 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.

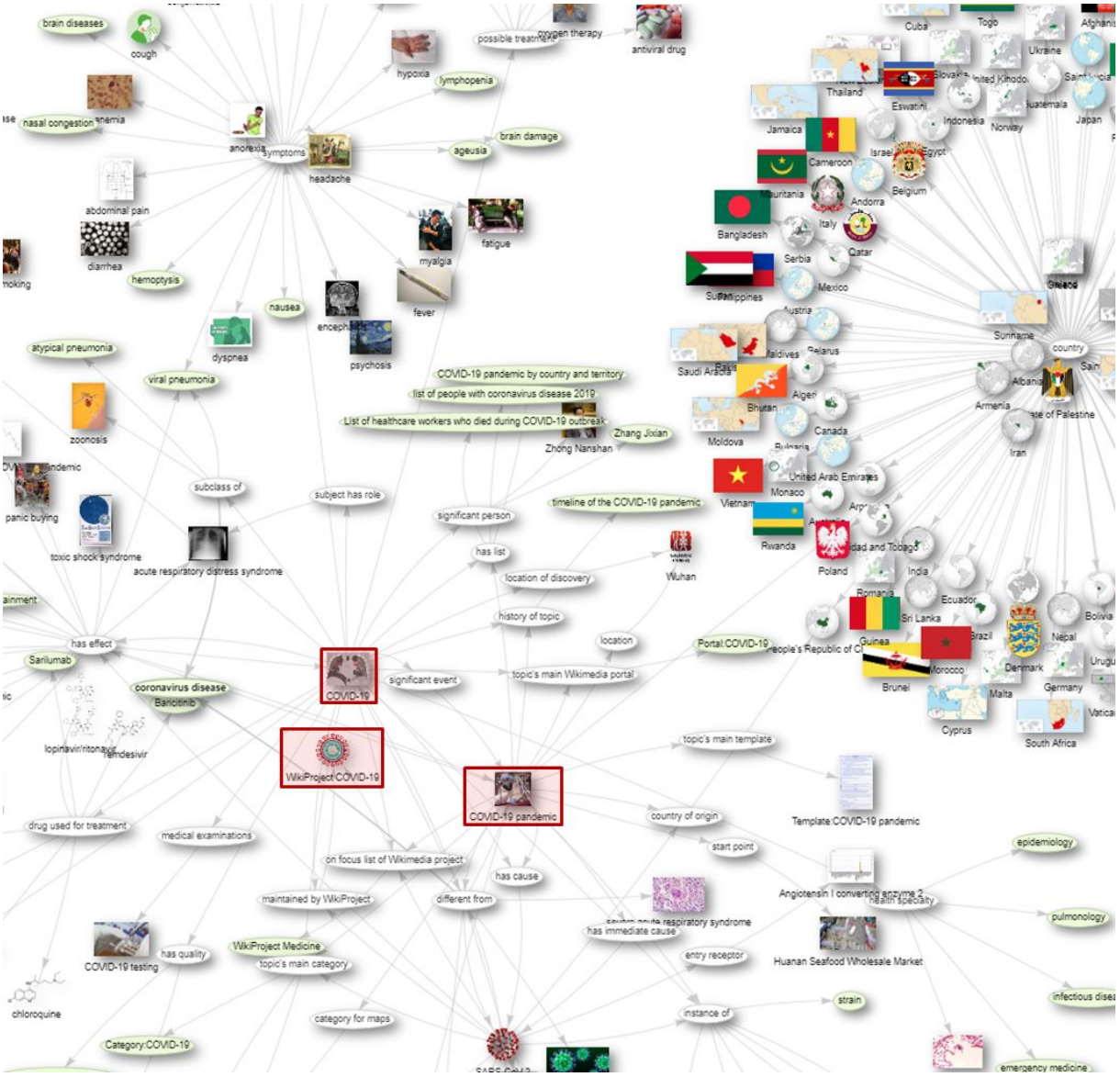


Figure S1: Snapshot of the extended graph of the three main COVID items and the statements for which they are the subject. Linked items demonstrate the variety of topics for which the three main COVID items (indicated in red) are the subject and present a small subset of the classes indicated in Fig. 2. (Available at: https://w.wiki/cPa, live data: https://tinyurl.com/y2ddm3n4, Access Date: August 19, 2020)

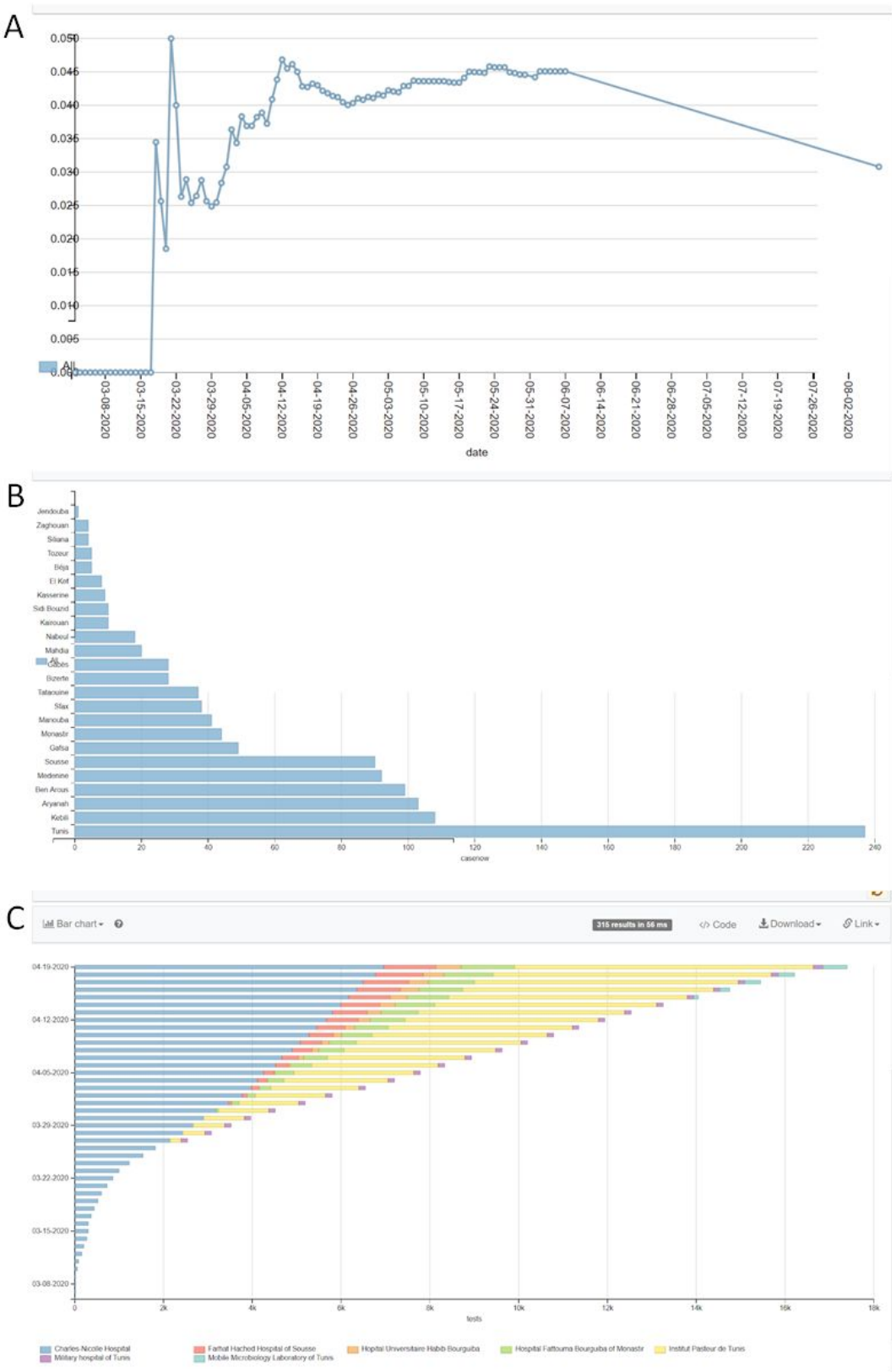


Figure S2. Epidemiological data for Tunisia (Available at: https://w.wiki/cQC). The SPEED website was set up as a COVID-19 data dashboard for Tunisia. A) Daily mortality rate from COVID-19 in Tunisia (live data: https://w.wiki/N2p). B) Tunisian governorate-level cases (live data: https://w.wiki/N9Y). 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 (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 (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). 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

Figure S6. Information regarding clinical trials on interventions to treat COVID-19 (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

Figure S7. Computer applications and their types (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

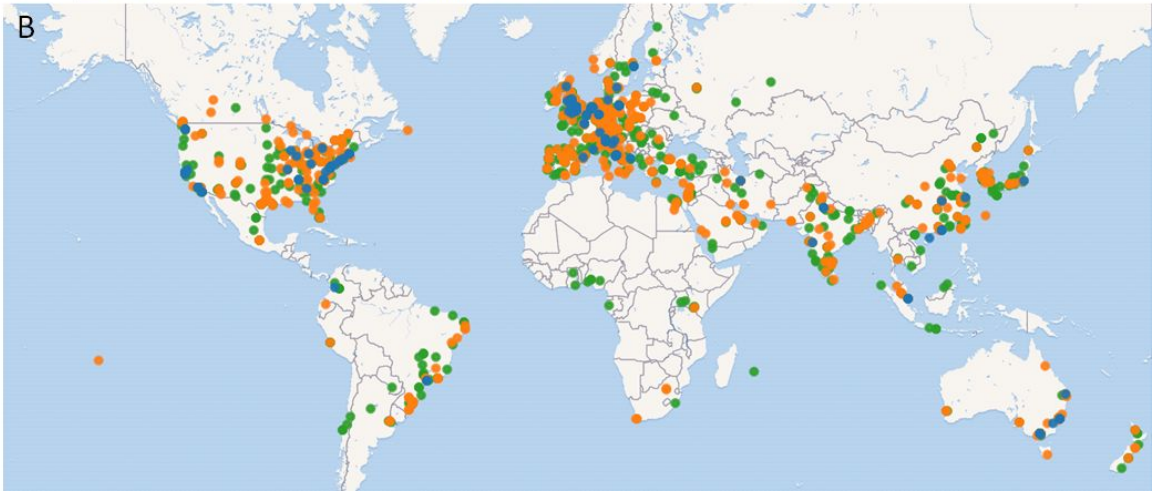


Figure S8. Information on authors of articles on COVID-related topics (Available at: https://w.wiki/cQh). A) Awards most frequently received by authors of C3 papers (live data: https://w.wiki/ban), B) Map of organizations associated with works about C3 with institutions that

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

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

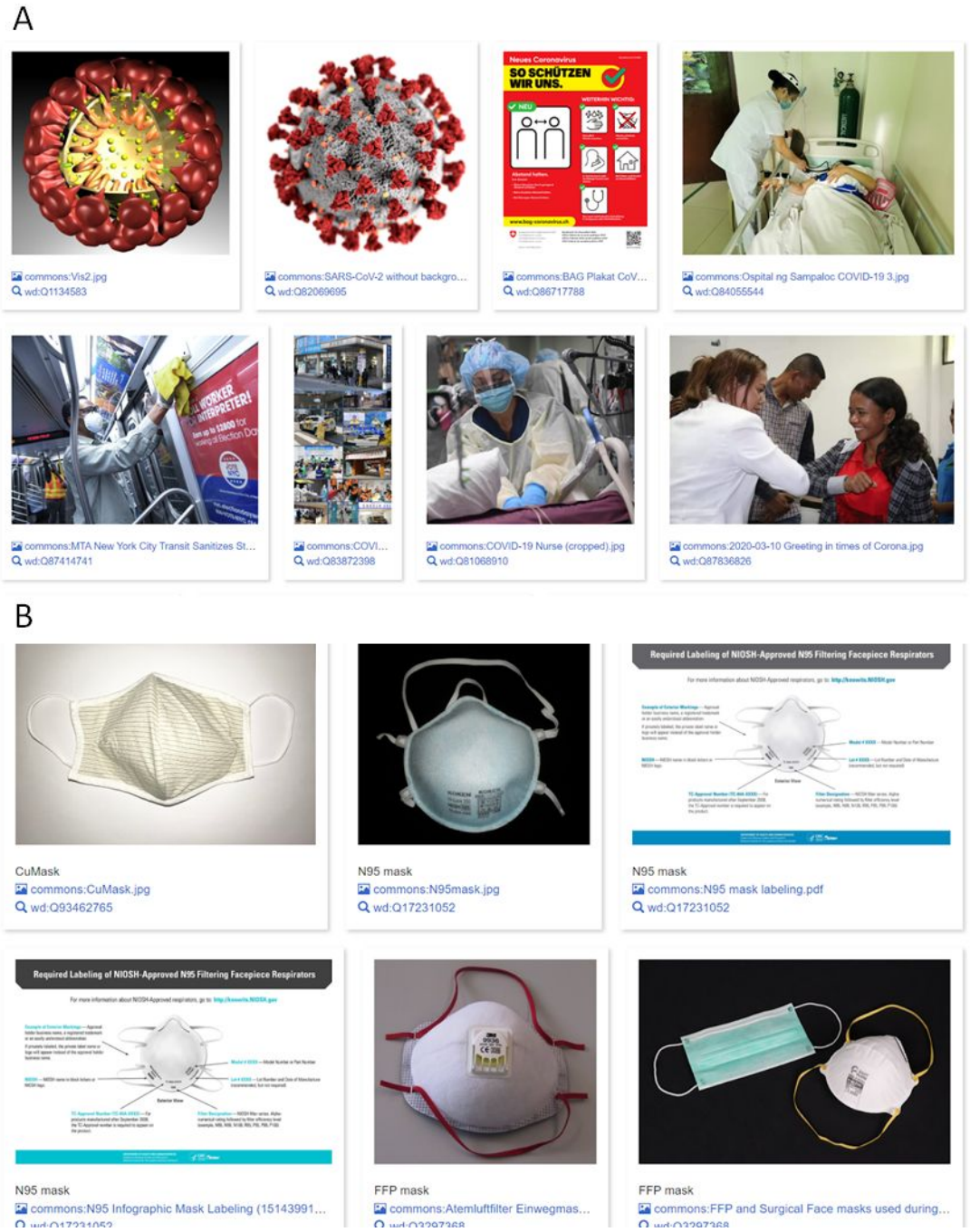


Figure S10. COVID-related images based on structured data (Available at: https://w.wiki/cOt). Images in wikimedia commons used to be organised solely by a hierarchical category structure. Since 2019, structured data can be associated with images via Wikidata statements. A) Images from Wikimedia Commons 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).