











METHODOLOGY ARTICLE

Open Access



# A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses

Andra Waagmeester<sup>1</sup> , Egon L. Willighagen<sup>2</sup> , Andrew I. Su<sup>3</sup> , Martina Kutmon<sup>2,4</sup> , Jose Emilio Labra Gayo<sup>5</sup> , Daniel Fernández-Álvarez<sup>5</sup> , Quentin Groom<sup>6</sup> , Peter J. Schaap<sup>7</sup> , Lisa M. Verhagen<sup>8</sup>  and Jasper J. Koehorst<sup>7\*</sup> 

## Abstract

**Background:** Pandemics, even more than other medical problems, require swift integration of knowledge. When caused by a new virus, understanding the underlying biology may help finding solutions. In a setting where there are a large number of loosely related projects and initiatives, we need common ground, also known as a “commons.” Wikidata, a public knowledge graph aligned with Wikipedia, is such a commons and uses unique identifiers to link knowledge in other knowledge bases. However, Wikidata may not always have the right schema for the urgent questions. In this paper, we address this problem by showing how a data schema required for the integration can be modeled with entity schemas represented by Shape Expressions.

**Results:** As a telling example, we describe the process of aligning resources on the genomes and proteomes of the SARS-CoV-2 virus and related viruses as well as how Shape Expressions can be defined for Wikidata to model the knowledge, helping others studying the SARS-CoV-2 pandemic. How this model can be used to make data between various resources interoperable is demonstrated by integrating data from NCBI (National Center for Biotechnology Information) Taxonomy, NCBI Genes, UniProt, and WikiPathways. Based on that model, a set of automated applications or bots were written for regular updates of these sources in Wikidata and added to a platform for automatically running these updates.

**Conclusions:** Although this workflow is developed and applied in the context of the COVID-19 pandemic, to demonstrate its broader applicability it was also applied to other human coronaviruses (MERS, SARS, human coronavirus NL63, human coronavirus 229E, human coronavirus HKU1, human coronavirus OC4).

**Keywords:** COVID-19, Wikidata, Linked data, ShEx, Open Science

## Background

The coronavirus disease 2019 (COVID-19) pandemic, caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) virus, is leading to a burst of swiftly released scientific publications on the matter

[1]. In response to the pandemic, many research groups have started projects to understand the SARS-CoV-2 virus life cycle and to find solutions. Examples of the numerous projects include [outbreak.info](#) [2], Virus Outbreak Data Network (VODAN) [3], [CORD-19-on-FHIR](#) [4], [KG-COVID-19 knowledge graph](#) [5], and the [COVID-19 Disease Map](#) [6]. Many research papers and preprints get published every week and many call for more Open Science [7]. The Dutch universities went a

\* Correspondence: [jasper.koehorst@wur.nl](mailto:jasper.koehorst@wur.nl)

<sup>7</sup>Department of Agrotechnology and Food Sciences, Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, The Netherlands

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

step further and want to make any previously published research openly available, in whatever way related to COVID-19 research.

However, this swift release of research findings comes with an increased number of incorrect interpretations [8] which can be problematic when new research articles are picked up by main-stream media [9]. Rapid evaluation of these new research findings and integration with existing resources requires frictionless access to the underlying research data upon which the findings are based. This requires interoperable data and sophisticated integration of these resources. Part of this integration is reconciliation, which is the process where matching concepts in Wikidata are sought [10]. Is a particular gene or protein already described in Wikidata? Using a shared interoperability layer, like Wikidata, different resources can be more easily linked.

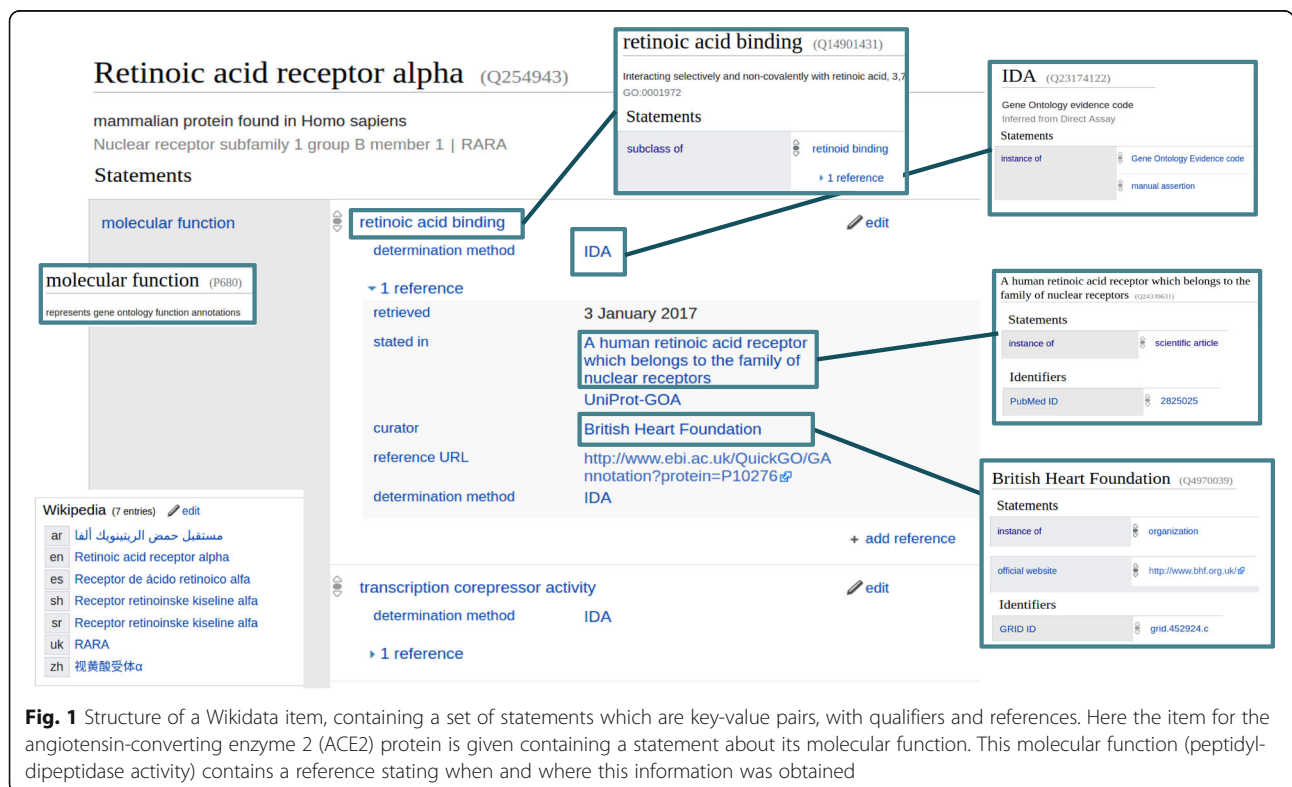
Wikidata is the linked-data repository of the Wikimedia Foundation. It is developed to provide Wikipedia and its sister projects with structured data. One interesting feature of Wikidata is that provenance and attribution can easily be included using the references and qualifiers which are core to the Wikidata data model.

The Gene Wiki project has been leveraging Wikidata to link different research silos by creating a brokerage system between resources on genetics, biological processes, related diseases, and associated drugs [11]. Various use cases ranging from crowdsourced curation of

biomedical ontologies, phenotype-based diagnosis of diseases, and drug repurposing can feed on this system. The project recognizes Wikidata as a sustainable infrastructure for scientific knowledge in the life sciences.

In contrast to legacy databases, where data models follow a relational data schema of connected tables, Wikidata uses statements to store facts (see Fig. 1) [11–14]. This model of statements aligns well with the RDF triple model of the semantic web and the content of Wikidata is also serialized as Resource Description Framework (RDF) triples [15, 16], acting as a stepping stone for data resources to the semantic web. Through its SPARQL (SPARQL Protocol and RDF Query Language) endpoint [17], knowledge captured in Wikidata can be integrated with other nodes in the semantic web, using mappings between these resources or through federated SPARQL queries [18]. A Wikidata item, as depicted in Fig. 1, has properties and values. The values are editable by everyone, but the property types are restricted. Creating new properties requires a property proposal that is formally discussed online. When there is consensus on the usefulness of the proposed property, it is created by a system administrator with its attached semantics. Users cannot create new properties at will, which makes it (together with its community acceptance) highly sustainable.

The Gene Wiki project aligns novel primary data sources with Wikidata in a cycle of consecutive steps



**Fig. 1** Structure of a Wikidata item, containing a set of statements which are key-value pairs, with qualifiers and references. Here the item for the angiotensin-converting enzyme 2 (ACE2) protein is given containing a statement about its molecular function. This molecular function (peptidyl-peptidase activity) contains a reference stating when and where this information was obtained

where the data schema of the primary source is aligned with the available Wikidata properties. Once the schema is in place, bots are developed to add and regularly update Wikidata with knowledge from the primary resource under scrutiny.

Automated editing of Wikidata simplifies the process; however, the quality control must be monitored carefully. This requires a clear data schema that allows the various resources to be linked together with their provenance. This schema describes the key concepts required for the integrations of the resources we are interested in the NCBI Taxonomy [19], NCBI Gene [20], UniProt [21], the Protein Data Bank (PDB) [22], WikiPathways [23], and PubMed [24]. Therefore, the core elements for which we need a model include viruses, virus strains, virus genes, and virus proteins. The first two provide the link to taxonomies, the models for genes and proteins link to UniProt, PDB, and WikiPathways. These key concepts are also required to annotate research output such as journal articles and datasets related to these topics. Wikidata calls such keywords “main subjects.” The introduction of this model and the actual SARS-CoV-2 genes and proteins in Wikidata enables the integration of these resources. The resources used were selected based on their eligibility for inclusion in Wikidata. Wikidata is available under a CC0 1.0 Universal (CC0 1.0) Public Domain Dedication which stipulates public use of the data included. Some valid resources use more restrictive licenses which prevents their inclusion in Wikidata.

This paper is a case report of a workflow/protocol for data integration and publication. The first step in this approach is to develop the data schema. Wikidata has a schema extension called EntitySchema that uses Shape Expressions (ShEx) as the structural schema language to describe and capture schemas of concepts [25, 26]. With ShEx, we describe the RDF structure by which Wikidata content is made available. These Shapes have the advantage that they are easily exchanged and describe linked-data models as a single knowledge graph. Since the Shapes describe the model, they enable discussion, revealing inconsistencies between resources, and allow for consistency checks of the content added by automated procedures. Eventually, we would like to get to a workflow where issues that can be fixed automatically are corrected, whereas biological inconsistencies will be made available for evaluation by field experts and non-domain specific issues are acted upon by the Wikidata community at large. With the model defined, the focus can turn to the process of adding knowledge to Wikidata. In this phase, the seven human coronaviruses (HCoVs), Middle East respiratory syndrome (MERS), SARS, SARS-CoV-2

(causing COVID-19), human coronavirus NL63, human coronavirus 229E, human coronavirus HKU1, and human coronavirus OC4 [27], can be added to Wikidata. This protocol is finalized by describing how the resulting data schema and data can be applied to support other projects, particularly the WikiPathways COVID Portal.

The Semantic Web was proposed as a vision of the Web, in which information is given well-defined meaning and better-enabling computers and people to work in cooperation [28]. In order to achieve that goal, several technologies have appeared, like RDF for describing resources [16], SPARQL to query RDF data [29], and the Web Ontology Language (OWL) to represent ontologies [30].

Linked data was later proposed as a set of best practices to share and reuse data on the web [31]. The linked data principles can be summarized in four rules that promote the use of uniform resource identifiers (URIs) to name things, which can be looked up to retrieve useful information for humans and for machines using RDF, as well as having links to related resources. These principles have been adopted by several projects, enabling a web of reusable data, known as the linked data cloud [32], which has also been applied to life science [33].

One prominent project is Wikidata, which has become one of the largest collections of open data on the web [18]. Wikidata follows the linked data principles offering both HTML and RDF views of every item with their corresponding links to related items, and a SPARQL endpoint called the Wikidata Query Service.

Wikidata’s RDF model offers a reification mechanism which enables the representation of information about statements like qualifiers and references [34]. For each statement in Wikidata, there is a direct property in the `wdt` namespace that indicates the direct value. In addition, the Wikidata data model adds other statements for reification purposes that allow enrichment of the declarations with references and qualifiers (for a topical treatise, see Ref. [35]). As an example, item Q14875321, which represents ACE2 (protein-coding gene in the *Homo sapiens* species) has a statement specifying that it can be found on chromosome (P1057) with value chromosome X (Q29867336). In RDF Turtle, this can be declared as:

```
wd:Q14875321 wdt:P1057 wd:Q29867336 .
```

That statement can be reified to add qualifiers and references. For example, a qualifier can state that the genomic assembly (P659) is GRCh38 (Q20966585) with a reference declaring that it was stated (P248) in Ensembl Release 99 (Q83867711). In Turtle, those declarations are represented as (see also Fig. 2):

```

wd:Q14875321 rdfs:label "ACE2"@en ;
wdt:P1057 wd:Q29867336 .

wd:Q14875321 p:P1057 [
ps:P1057 wd:Q29867336 ;
pq:P659 wd:Q20966585 ;
prov:wasDerivedFrom [
pr:P248 wd:Q83867711
] ] .

wd:Q29867336 rdfs:label "human X chromosome" .
wd:Q20966585 rdfs:label "Genome assembly GRCh38" .
wd:Q83867711 rdfs:label "ensembl Release 99" .
    
```

**Results**

**Semantic data landscape**

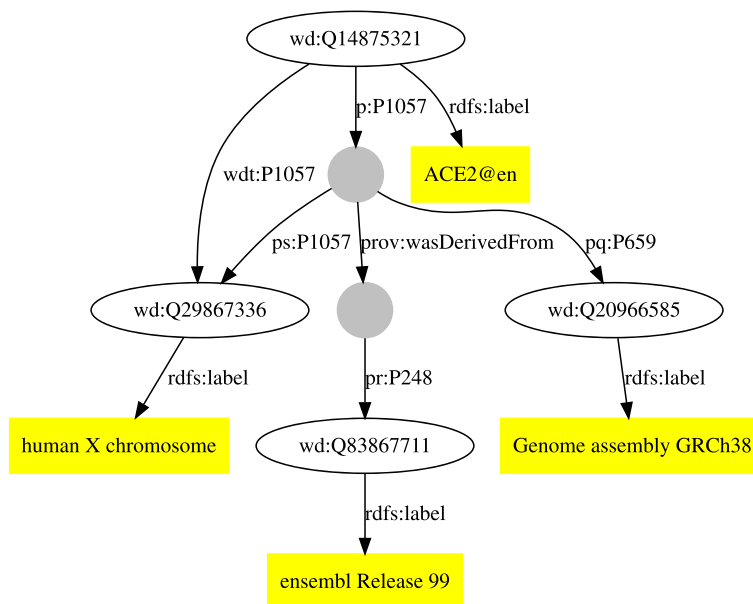
To align the different sources in Wikidata, a common data schema is needed. We have created a collection of schemas that represent the structure of the items added to Wikidata. Input to the workflow is the NCBI taxon identifier, which is input to [mygene.info](#) (see Fig. 8b). Taxon information is obtained and added to Wikidata according to a set of linked Entity Schemas (virus, strain, disease) [36–39]. Gene annotations are obtained and added to Wikidata following the schemas virus gene and protein annotations [40, 41] are obtained and added to Wikidata following the two schemas. Pathway information is sourced according to the schema describing Wikipathways representation in Wikidata [42]. The last two schemas are an extension from more generic schemas for proteins [42] and genes [43] (Fig. 3).

**ShEx validation**

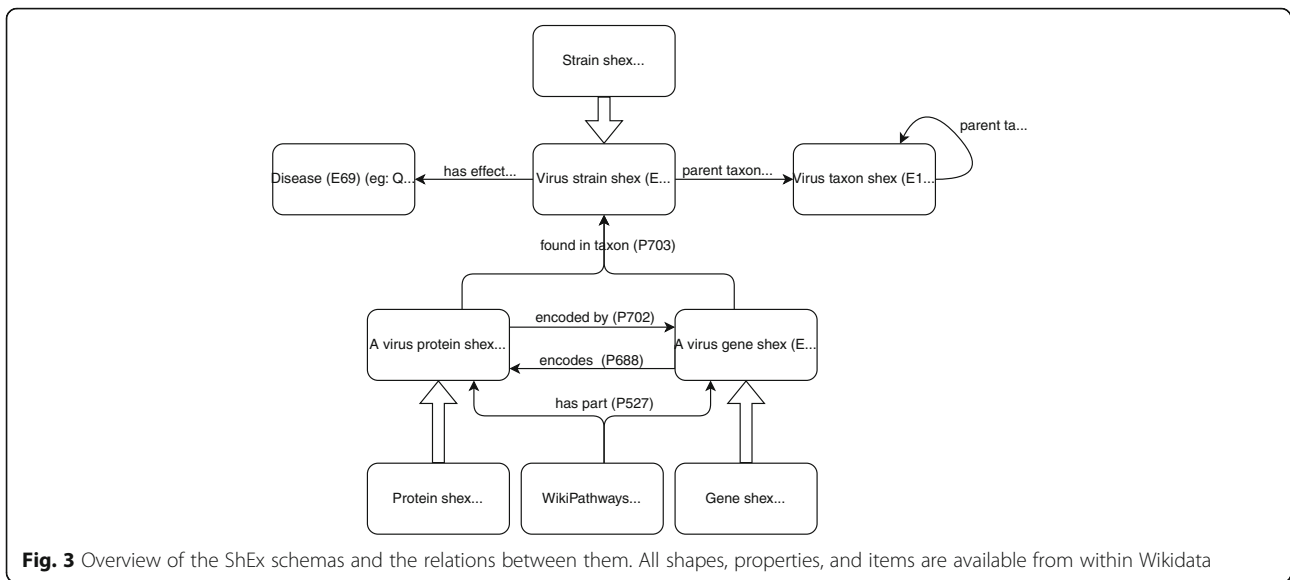
With the set of ShEx schemas, it is possible to check if existing data aligns with the expressed expectations. Figure 9 demonstrates two cases in which one Wikidata item (Q70644554) does not align with the tested Schema E174, while another Wikidata item on a Wikipathways Pathway (Q88292589) does conform to schema E41. The Wikidata EntitySchema extension does allow checking for conformance. There are currently five actively maintained ShEx implementations that allow checking for ShEx conformance at a larger scale [25] (Fig. 4).

**Bots**

The bots developed and used in this protocol are adaptations of the bots developed in the Gene Wiki project. On regular intervals, the bots run to update viral gene and protein annotations as well as pathway updates from WikiPathways. For the gene and protein annotations, we



**Fig. 2** Example of an RDF data model representing ACE2, created with RDFShape [32]



**Fig. 3** Overview of the ShEx schemas and the relations between them. All shapes, properties, and items are available from within Wikidata

have also made a Jupyter Notebook. The bot that synchronizes virus, gene, and protein information and the Jupyter Notebook are available [44]. The bot that synchronizes the WikiPathways pathways with Wikidata was updated from the original version to allow adding proteins annotated with Wikidata identifiers and no longer requires pathways to be part of the WikiPathways *Curated Collection*. The customized bot source code is available [45].

Both bots are now part of the automation server used in the Gene Wiki project. This runs on the

Jenkins platform [46]. Although Jenkins is mainly aimed at software deployments, its extended scheduling capabilities the synchronization procedure at set intervals that can be changed depending on the update speed of the external resources. The Jenkins jobs are also available from [47].

**Data added**

Using the gene and proteins bots explained in the “Methods” section, missing genes and proteins have

**Panel A:** Query: Entities to check. SELECT ?item WHERE ( ?item wdt:P31 wd:Q855769. ) LIMIT 1. Result: ST8:USA300 (Q150908) is a strain. Statements: instance of strain (+ 1 reference).

**Panel B:** Query: Entities to check. SELECT ?pathway WHERE (VALUES ?pathway (wd:Q88292589) ?pathway ?p ?o .). Result: SARS-CoV-2 and COVID-19 Pathway (Q88292589). Description: Collaborative project for curation biological processes involved in the COVID-19 disease after SARS-CoV-2 infection. Statements: instance of biological pathway (+ 1 reference).

**Fig. 4** Application of the drafted ShEx schemas in the EntitySchema extension of Wikidata allows for confirmation if a set of on-topic items align with expressed expectations. In panel a, the application renders the Wikidata item invalid due to a missing reference which in turn does not conform to the expressed ShEx whereas in panel b, the item (Q88292589) conforms to the applied schema

been added for the seven human coronaviruses. The results are summarized in Table 1. The automatically added and updated gene and protein items were manually curated. For SARS-CoV-2, all items were already manually created, and the bot only edited gene items. Thirteen out of 27 protein entries were created by the authors. For the other species, all gene entries and most protein entries have been created by the bot. Only for MERS and SARS-CoV-2, some protein entries were added manually, including some by us. During this effort, which took 3 weeks, the bot created a number of duplicates. These have been manually corrected. It should also be noted that for SARS-CoV-2 many proteins and protein fragments do not have RefSeq or UniProt identifiers, mostly for the virus protein fragments.

**Use cases**

**BridgeDb**

Using the dedicated code to create a BridgeDb identifier mapping database for coronaviruses, mappings were extracted from Wikidata with a SPARQL query for the seven human coronaviruses and the SARS-related viruses. This resulted in a mapping database with 567 mappings between 380 identifiers (version 2020-11-30). This includes 171 Wikidata identifiers, 70 NCBI Gene identifiers, 71 UniProt identifiers, 58 RefSeq identifiers, and 10 Guide to Pharmacology Target identifiers. The mapping file has been released on the BridgeDb website and archived on Zenodo [48]. The mapping database has also been loaded on the BridgeDb webservice which means it can be used in the next use case: providing links out for WikiPathways.

**WikiPathways**

The WikiPathways project is involved in an international collaboration to curate knowledge about the biological processes around SARS-CoV-2 and COVID-19. The authors have started a pathway specifically about SARS-CoV-2 ([wikipathways:WP4846](https://www.wikipathways.org/wiki/WP4846)).

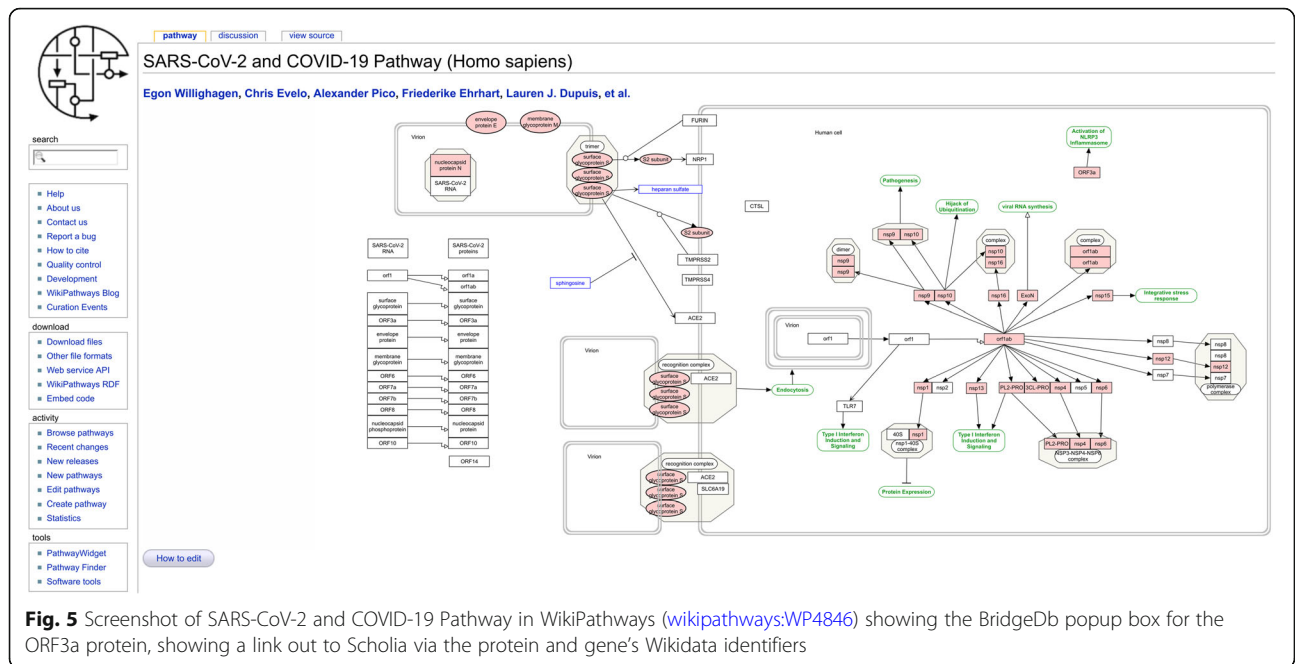
To ensure interoperability, WikiPathways uses BridgeDb and taking advantage of the enriched BridgeDb webservice, WikiPathways now links out for HCoV genes and proteins (depending on availability of mappings) to RefSeq, NCBI Gene, UniProt, and Scholia (see Fig. 5). The latter links to the next use case and provides a link to literature about the virus. It should be noted that for each gene and protein two Wikidata identifiers with links may be given. In that case, one is for the gene and one for the protein.

**Scholia**

The WikiPathways use case shows us that literature describes our knowledge about how coronaviruses work at a rather detailed level. Indeed, many articles discuss the genetics, homology of genes and proteins across viruses, or the molecular aspects of how these proteins are created and how they interfere with the biology of the human cell. The biological pathways show these processes, but ultimately the knowledge comes from literature. Wikidata allows us to link literature to specific virus proteins and genes, depending on what the article describes. For this, it uses the “main subject” (P921) property [49]. We manually annotated literature with the Wikidata items for specific proteins and genes, particularly useful for virus concepts for which reference databases do not provide entries, such as the non-structural proteins. We developed two SPARQL queries to count the number of links between genes [50] and proteins [51] and the articles that discuss them. Scholia takes advantage of the “main subject” annotation, allowing the creation of “topic” pages for each protein. For example, Fig. 6 shows the topic page of the SARS-CoV-2 spike protein. As such, Scholia provides a simple user interface summarizing literature about a specific feature of the SARS-CoV-2 virus. An RSS feed is even available to get alerted about new literature about each topic (also visible in Fig. 6).

**Table 1** Summary of the seven human coronaviruses, including taxon identifiers, the Wikidata items, and the number of genes and proteins. The latter two are generated by the SPARQL queries `geneCount.rq` and `proteinCount.rq` in Additional file 1

Virus strain	NCBI Taxon ID	Wikidata Qid	# Genes	# Proteins
SARS virus	694009	Q278567	14	11
Middle East respiratory syndrome coronavirus	1335626	Q4902157	11	9
Human coronavirus NL63	277944	Q8351095	7	6
Human coronavirus 229E	11137	Q16983356	8	8
Human coronavirus HKU1	290028	Q16983360	9	9
Human coronavirus OC43	31631	Q16991954	9	8
SARS-CoV-2	2697049	Q82069695	11	27



**Fig. 5** Screenshot of SARS-CoV-2 and COVID-19 Pathway in WikiPathways ([wikipathways:WP4846](https://www.wikipathways.org/WP4846)) showing the BridgeDb popup box for the ORF3a protein, showing a link out to Scholia via the protein and gene's Wikidata identifiers

**Discussion**

This paper describes a protocol we developed to align genetic annotations from reference resources with Wikidata. Numerous annotations are scattered across different sources without any overall integration, thereby reducing the reusability of knowledge from different sources. Integration of the annotations from these resources is a complex and time-consuming task. Each

resource uses different ways to access the data from a user and machine perspective. Making use of these protocols programmatically to access and retrieve data of interest requires the knowledge of various technologies and procedures to extract the information of interest.

Wikidata provides a solution. It is part of the semantic web, taking advantage of its reification of the Wikidata items as RDF. Data in Wikidata itself is frequently, often



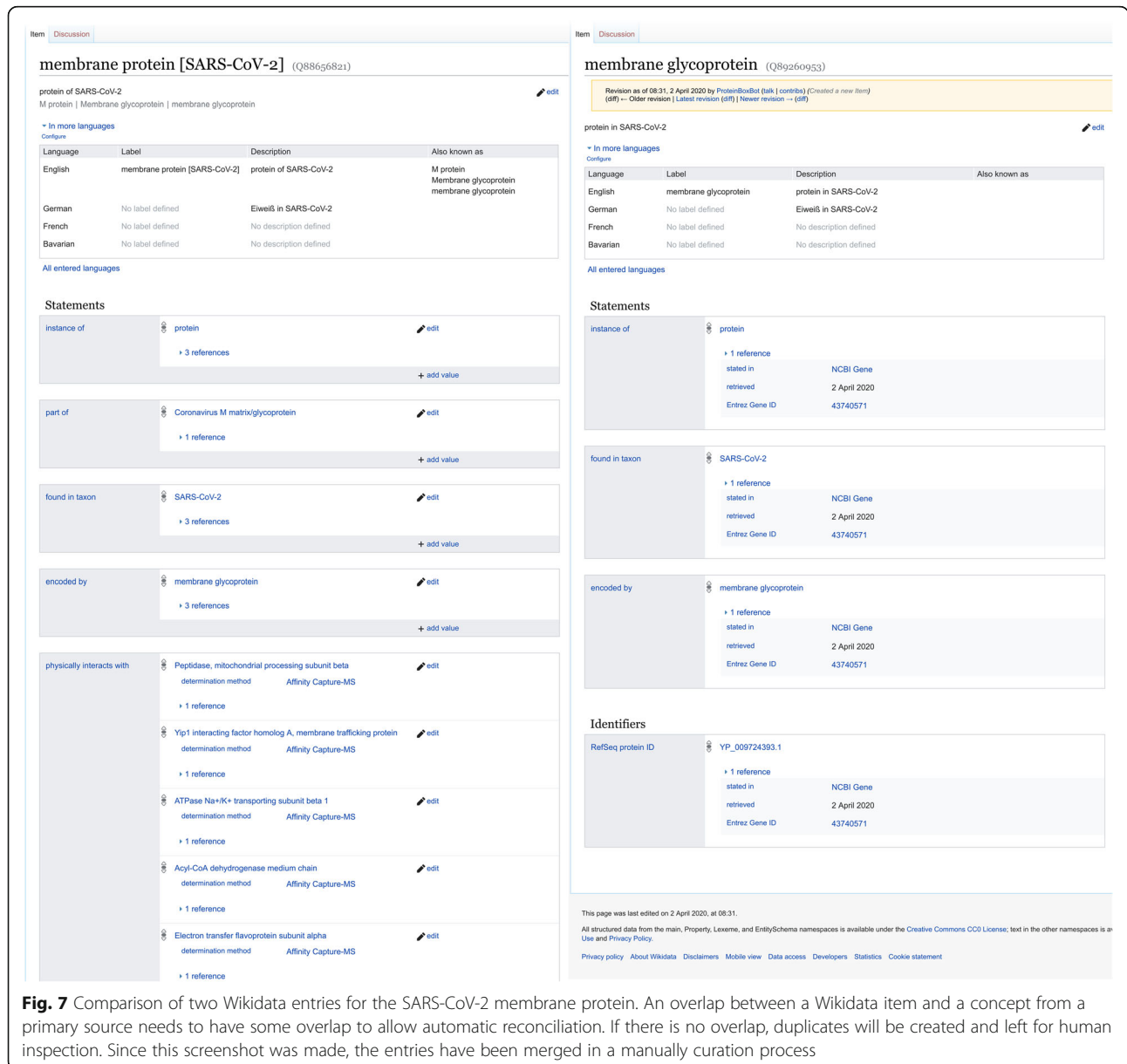
**Fig. 6** Screenshot of the Scholia page for the SARS-CoV-2 spike glycoprotein, it shows four articles that specifically discuss this protein

almost instantaneously, synchronized with the RDF resource and available through its SPARQL endpoint [17]. The modeling process turns out to be an important aspect of this protocol. Wikidata contains numerous entity classes as entities and more than 7000 properties which are ready for (re-)use. However, that also means that this is a confusing landscape to navigate. The ShEx Schema has helped us develop a clear model. This is a social contract between the authors of this paper, as well as documentation for future users.

Using these schemas, it was simpler to validate the correctness of the updated bots to enter data in Wikidata. The bots have been transferred to the Gene Wiki Jenkins platform. This allows the bots to be kept

running regularly, pending on the ongoing efforts of the coronavirus and COVID-19 research communities. While the work of the bots will continue to need human oversight, potentially to correct errors, it provides a level of scalability and generally alleviates the authors from a lot of repetitive work.

One of the risks of using bots is the possible generation of duplicate items. Though this is also a risk in manual addition of items, humans can apply a wider range of academic knowledge to resolve these issues. Indeed, in running the bots, duplicate Wikidata items were created, for which an example is shown in Fig. 7. The Wikidataintegrator library does have functionality to prevent the creation of duplicates by comparing



**Fig. 7** Comparison of two Wikidata entries for the SARS-CoV-2 membrane protein. An overlap between a Wikidata item and a concept from a primary source needs to have some overlap to allow automatic reconciliation. If there is no overlap, duplicates will be created and left for human inspection. Since this screenshot was made, the entries have been merged in a manually curation process



properties, based on used database identifiers. However, if two items have been created using different identifiers, these cannot be easily identified.

Close inspection of examples, such as the one in Fig. 7, showed that the duplicates were created because there was a lack of overlap between the data to be added and the existing item. The UniProt identifier did not yet resolve, because it was manually extracted from information in the March 27 pre-release (but now part of the regular releases). In this example, the Pfam protein families database [52] identifier was the only identifier upon which reconciliation could happen. However, that identifier pointed to a webpage that did not contain mappings to other identifiers. In addition, the lack of references to the primary source hampers the curator's ability to merge duplicate items and expert knowledge was essential to identify the duplication. Fortunately, names used for these RNA viruses only refer to one protein as the membrane protein. Generally, the curator would have to revert to the primary literature to identify the overlap. Statements about "encoded by" to the protein coding genes were found to be helpful as well. Reconciliation might be possible through sequence alignment, which means substantial expert knowledge and skills are required.

This makes reconciliation in Wikidata based on matching labels, descriptions and synonyms, matching statements, and captured provenance (qualifiers and references) hazardous, due to different meanings to the same label. A targeted curation query (*geneAndProtein-CurationQuery.rq*, see Additional file 1) was developed to highlight such duplications and manually curated seven duplicate protein entries for SARS-CoV-2 alone. This duplication is common and to be expected, particularly in rapidly evolving situations like a pandemic, when many groups contribute to the same effort. In this case, this paper only represents one group contributing to the *Wikidata:WikiProject COVID-19* [53].

We also discovered that virus taxonomy is different from zoological taxonomy. For example, there is no clear NCBI taxon identifier for SARS-CoV-1 and after consultation with other projects, we defaulted to using the taxon identifier for the SARS-related CoVs, something that NCBI and UniProt seem to have done as well.

Finally, we note that during the 2 weeks this effort took place, several other resources introduced overviews, including dedicated COVID-19 portals from UniProt [54] and the Protein DataBank in Europe [55].

## Conclusion

This manuscript presents a protocol to link information from disparate resources, including NCBI Taxonomy, NCBI Gene, UniProt, PubMed, and WikiPathways. Using the existing Wikidata infrastructure, we developed

semantic schemas for virus strains, genes, and proteins; bots written in Python to add knowledge on genes and proteins of the seven human coronaviruses and linked them to biological pathways in WikiPathways and to primary literature, visualized in Scholia. We were able to do so in the period of 2 weeks, using an ad hoc team from existing collaborations, taking advantage of the open nature of the components involved.

## Methods

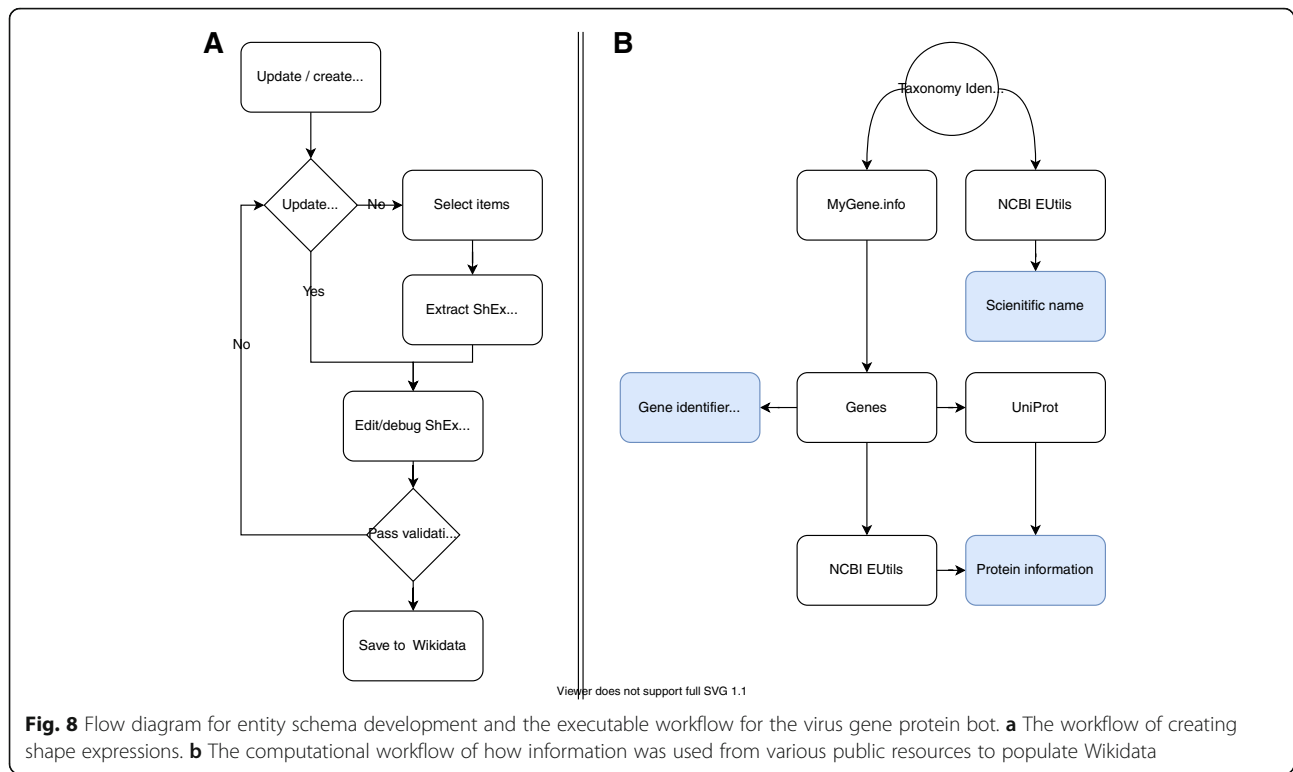
### Specifying data models with ShEx

Although the RDF data model is flexible, specifying an agreed structure for the data allows domain experts to identify the properties and structure of their data facilitating the integration between heterogeneous data sources. Shape Expressions were used to provide a suitable level of abstraction. Yet Another ShEx Editor (YaShE) [56], a ShEx editor implemented in JavaScript, was applied to author these Shapes [57]. This application provides the means to associate labels in the natural language of Wikidata to the corresponding identifiers. The initial entity schema was defined with YaShE as a proof of concept for virus genes and proteins. In parallel, statements already available in Wikidata were used to automatically generate an initial shape for virus strains with sheXer [58]. The statements for virus strains were retrieved with SPARQL from the Wikidata Query Service (WDQS) [17]. The generated Shape was then further improved through manual curation. The syntax of the Shape Expressions was continuously validated through YaShE and the Wikidata Entity Schema namespace was used to share and collaboratively update the schema with new properties. Figure 8 gives a visual outline of these steps.

### Populating Wikidata with human coronavirus data

The second step in our workflow is to add entries for all virus strains, genes, and their gene products to Wikidata. This information is spread over different resources. Here, annotations were obtained from NCBI EUtils [59], *Mygene.info* [60], and UniProt [21], as outlined below. Input to the workflow is the NCBI Taxonomy identifier of a virus under scrutiny. (e.g., 2697049 for SARS-CoV-2). The taxon annotations are extracted from NCBI EUtils. The gene and gene product annotations are extracted from *mygene.info* and the protein annotations are extracted from UniProt using the SPARQL endpoint [61].

Genomic information from seven human coronaviruses (HCoV) was collected, including the NCBI Taxonomy identifiers. For six virus strains, a reference genome was available and was used to populate Wikidata. For SARS-CoV-1, the NCBI Taxonomy identifier referred to various strains, though no reference strain



was available. To overcome this issue, the species taxon for SARS-related coronaviruses (SARSr-CoV) was used instead, following the practices of NCBI Genes and UniProt (Fig. 9).

**NCBI Eutils**

The Entrez Programming Utilities (EUtils) [59] is the application programming interface (API) to the Entrez query and database system at the NCBI. From this set of services the scientific name of the virus under scrutiny was extracted (e.g., “Severe acute respiratory syndrome coronavirus 2”).

**Mygene.info**

Mygene.info [60] is a web service that provides a REST API that can be used to obtain up-to-date gene annotations. The first step in the process is to get a list of applicable genes for a given virus by providing the NCBI taxon id. The following step is to obtain gene annotations for the individual genes from mygene.info (e.g., [62]). This results in the name and a set of applicable identifiers (Fig. 10).

**UniProt**

The annotations retrieved from mygene.info also contain protein identifiers such as UniProt [21], RefSeq [63], and PDB [22]; however, their respective names are lacking. To obtain names and mappings to other protein

identifiers, RefSeq and UniProt were consulted. Refseq annotations were acquired using the earlier mentioned NCBI EUtils. UniProt identifiers are acquired using the SPARQL endpoint of UniProt, which is a rich resource for protein annotations provided by the Swiss Bioinformatics Institute. Figure 5 shows the SPARQL query that was applied to acquire the protein annotations.

**Reconciliation with Wikidata**

Before the aggregated information on viruses, genes, and proteins can be added to Wikidata, reconciliation with Wikidata is necessary. If Wikidata items exist, they are updated; otherwise, new items are created. Reconciliation is driven by mapping existing identifiers in both the primary resources and Wikidata. It is possible to reconcile based on strings, but this is dangerous due to the ambiguity of the labels used [64]. When items on concepts are added to Wikidata that lack identifiers overlapping with the primary resource, reconciliation is challenging. Based on the Shape Expressions, the following properties are identified for use in reconciliation. For proteins, these are Uniprot ID (P352) and RefSeq protein ID (P637). For genes, these are NCBI Gene ID (P351) and Ensembl Gene (ID) (P594). When sourced information matches none of these properties, then a new item is created, if the concepts from the primary source reconcile with Wikidata items these are updated.

```
{
  "_id": "43740571",
  "_score": 15.594226,
  "accession": {
    "genomic": [
      "MN908947.3",
      "NC_045512.2"
    ],
    "protein": [
      "QHD43419.1",
      "YP_009724393.1"
    ]
  },
  "entrezgene": "43740571",
  "locus_tag": "GU280_gp05",
  "name": "membrane glycoprotein",
  "other_names": "membrane glycoprotein",
  "refseq": {
    "genomic": "NC_045512.2",
    "protein": "YP_009724393.1"
  },
  "retired": 43560233,
  "symbol": "M",
  "taxid": 2697049,
  "type_of_gene": "protein-coding"
}
```

**Fig. 9** JavaScript Object notation output of the [mygene.info](https://mygene.info) output for gene with NCBI gene identifier 43740571

**Wikidataintegrator** Wikidata integrator is a Python library [65] that wraps around the [Wikidata API](https://wikidata-api.readthedocs.io/) [11, 66]. From external resources such as the NCBI, gene and taxonomy statements have been compiled with provenance and assigned to the associated Wikidata items. When an item did not exist (or was not recognized), it was created. The module compiled a list of statements by parsing the primary sources under scrutiny and extracted what statements already existed on Wikidata. A JavaScript Object Notation (JSON) string was created that resembled the JSON data model used by the Wikidata API. This JSON string was then submitted to the Wikidata API for ingestion.

#### Data integration use cases/validation

##### *WikiPathways and BridgeDb*

WikiPathways is a biological pathway database and can visualize the details of interactions between genes, proteins, metabolites, and other entities participating in biological processes. It depends on BridgeDb to map identifiers of external data and knowledge to the identifiers used for the genes, proteins, and metabolites in the pathways [67]. Furthermore, mappings to Wikidata are required to establish the link of biological entities in pathways and journal articles that have those entities as main topics. Therefore, the virus genes and proteins are required to exist in Wikidata, enabling the interoperability

```

PREFIX uniprotkb: <http://purl.uniprot.org/uniprot/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?id ?label ?database
WHERE {
SERVICE <https://sparql.uniprot.org/sparql> {
VALUES ?database {
<http://purl.uniprot.org/database/PDB>
<http://purl.uniprot.org/database/RefSeq>
}
uniprotkb:uniprotID rdfs:label ?label ;
rdfs:seeAlso ?id .
?id <http://purl.uniprot.org/core/database> ?database .
}
}

```

**Fig. 10** The UniProt SPARQL query used to obtain additional protein annotations, descriptions, and external resources

between WikiPathways and Wikidata. Additionally, new virus mapping databases for BridgeDb are created by extracting the new virus gene and protein data, including links between Wikidata, NCBI Gene, RefSeq, UniProt, and Guide to Pharmacology Target identifiers using a SPARQL query [68]. The mapping databases will be updated regularly and will allow pathway curators to annotate virus genes and proteins in their pathways and provide link outs on the WikiPathways website.

The COVID-19-related pathways from WikiPathways COVID-19 Portal [9] are added to Wikidata using the approach previously described [11]. For this, a dedicated repository has been set up to hold the Graphical Pathway Markup Language (GPML) files, the internal WikiPathways file format [69]. The GPML is converted into RDF files with the WikiPathways RDF generator [23], while the files with author information are manually edited. For getting the most recent GPML files, a custom Bash script was developed (getPathways.sh in [70]). The conversion of the GPML to RDF uses the previously published tools for WikiPathways RDF [23]. Here, we adapted the code with a unit test that takes the pathways identifier as a parameter. This test is available in the SARS-CoV-2-WikiPathways branch of GPML2RDF along with a helper script (createTurtle.sh). Based on this earlier generated pathway RDF and using the Wikidataintegrator library, the WikiPathways bot was used to populate Wikidata with additional statements and items. The pathway bot was extended with the capability to link virus proteins to the corresponding pathways, which was essential to support the Wikidata resource. These changes can be found in the *sars-cov-2-wikipathways-2* branch.

### Scholia

The second use case is to demonstrate how we can link virus gene and protein information to literature. Here, we used Scholia [14] as a central tool. It provides a graphical interface

around data in Wikidata, for example, literature about a specific coronavirus protein (e.g., Q87917585 for the SARS-CoV-2 spike protein). Scholia uses SPARQL queries to provide information about topics. We annotated literature around the HCoV with the specific virus strains, the virus genes, and the virus proteins as “main topic.”

### Supplementary Information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12915-020-00940-y>.

**Additional file 1.** Queries.txt.

### Acknowledgements

The protocol presented in this work would not be as powerful if it was not for the works of many, which ensure that the fruits of our work are immediately interoperable with lots of other efforts, even though that integration was the whole purpose of developing this protocol. We thank the work of the larger Wikidata:WikiProject COVID-19 community in which we operated, Wikidata volunteers in general, and in particular, Tiago Lubiana, Ralf Stephan, Daniel Mietchen, and Finn Nielsen. Second, the work is partly driven by a collaboration in the WikiPathways COVID-19 project, and we particularly thank Anders Riutta for setting up a commit hook to the repository with WikiPathways COVID-19 pathways and the authors of these pathways.

### Authors' contributions

AW, EW, JK: conceptual design, code development, and wrote the manuscript. MK: code development and wrote the manuscript. JL, code development, visualization and wrote the manuscript. DF: minor review and code development. QG: virus-taxonomy review and wrote the manuscript. AS, PS, LM: review and editing. The authors read and approved the final manuscript.

### Funding

National Institute of General Medical Sciences (R01 GM089820)  
 • Andra Waagmeester, Andrew I. Su  
 Alfred P. Sloan Foundation (grant number G-2019-11458)  
 • Egon Willighagen  
 Spanish Ministry of Economy and Competitiveness (Society challenges: TIN2017-88877-R)  
 • Jose Emilio Labra Gayo, Daniel Fernández Álvarez  
 Netherlands Organisation for Scientific Research funded UNLOCK project (NRGWl.obrug.2018.005)  
 • Jasper Koehorst, Peter Schaap

SYNTHESIS+ a Research and Innovation action funded under H2020-EU.1.4.1.2. Grant agreement ID: 823827.

- Quentin Groom
- ZonMw (grant number: 10430012010015)
- Martina Kutmon, Egon Willighagen

**Availability of data and materials**

All data and corresponding schema are available in Wikidata. All source code is available from GitHub [36–44, 71].

**Ethics approval and consent to participate**

No human participants, human material, or human data were involved in the work leading to this manuscript.

**Consent for publication**

Not applicable.

**Competing interests**

All authors have declared to have no competing interests.

**Author details**

<sup>1</sup>Micelio, Antwerpen, Belgium. <sup>2</sup>Department of Bioinformatics – BiGCaT, NUTRIM, Maastricht University, Maastricht, The Netherlands. <sup>3</sup>Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA. <sup>4</sup>Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, The Netherlands. <sup>5</sup>WESO Research Group, University of Oviedo, Oviedo, Spain. <sup>6</sup>Meise Botanic Garden, Meise, Belgium. <sup>7</sup>Department of Agrotechnology and Food Sciences, Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, The Netherlands. <sup>8</sup>Intravacc, PO Box 450, 3720 AL Bilthoven, The Netherlands.

Received: 12 June 2020 Accepted: 13 December 2020

Published online: 22 January 2021

**References**

1. Watkins J. Preventing a covid-19 pandemic. *BMJ*. 2020;368. <https://doi.org/10.1136/bmj.m810>.
2. [outbreakinfo.org](https://outbreakinfo.org/). <https://outbreakinfo.org/>. Accessed 25 Nov 2020.
3. Virus Outbreak Data Network (VODAN). GO FAIR. <https://www.go-fair.org/implementation-networks/overview/vodan/>. Accessed 25 Nov 2020.
4. fhircat/CORD-19-on-FHIR. Python. FHIRCat; 2020. <https://github.com/fhircat/CORD-19-on-FHIR>. Accessed 25 Nov 2020.
5. Reese JT, Unni D, Callahan TJ, Cappelletti L, Ravanmehr V, Carbon S, et al. KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. *Patterns*. 2020;10015. <https://doi.org/10.1016/j.patter.2020.100155>.
6. Ostaszewski M, Mazein A, Gillespie ME, Kuperstein I, Niarakis A, Hermjakob H, et al. COVID-19 disease map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Sci Data*. 2020;7:136.
7. Coronavirus and Open Science: our reads and open use cases. SPARC Europe. <https://sparceurope.org/coronaopensciencereadsandusecases/>. Accessed 25 Nov 2020.
8. Sharma M, Scarr S, Kell K. Speed science: Reuters. <https://graphics.reuters.com/CHINA-HEALTH-RESEARCH/0100B5E53MG/index.html>. Accessed 25 Nov 2020.
9. Mahase E. Covid-19: six million doses of hydroxychloroquine donated to US despite lack of evidence. *BMJ*. 2020;368:m1166.
10. Wikidata. [https://www.wikidata.org/wiki/Wikidata:Main\\_Page](https://www.wikidata.org/wiki/Wikidata:Main_Page). Accessed 25 Nov 2020.
11. Waagmeester A, Stupp G, Burgstaller-Muehlbacher S, Good BM, Griffith M, Griffith OL, et al. Wikidata as a knowledge graph for the life sciences. *eLife*. 2020;9:e52614.
12. Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase. *Commun ACM*. 2014;57:78–85.
13. Burgstaller-Muehlbacher S, Waagmeester A, Mittra E, Turner J, Putman T, Leong J, et al. Wikidata as a semantic framework for the Gene Wiki initiative. *Database J Biol Databases Curation*. 2016;2016. <https://doi.org/10.1093/database/baw015>.
14. Nielsen FÅ, Mietchen D, Willighagen E, Scholia, Scientometrics and Wikidata. In: Blomqvist E, Hose K, Paulheim H, Ławrynowicz A, Ciravegna F, Hartig O,

- editors. *The semantic web: ESWC 2017 satellite events*. Cham: Springer International Publishing; 2017. p. 237–59.
15. Exleben F, Günther M, Krötzsch M, Mendez J, Vrandečić D. Introducing Wikidata to the Linked Data Web. In: Mika P, Tudorache T, Bernstein A, Welty C, Knoblock C, Vrandečić D, et al., editors. *The Semantic Web – ISWC 2014*. Cham: Springer International Publishing; 2014. p. 50–65.
16. RDF 1.1 Concepts and Abstract Syntax. <https://www.w3.org/TR/rdf11-concepts/>. Accessed 25 Nov 2020.
17. Wikidata Query Service. <https://query.wikidata.org/>. Accessed 25 Nov 2020.
18. Getting the Most out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph - International Center for Computational Logic. <https://iccl.inf.tu-dresden.de/web/Inproceedings3044/en>. Accessed 25 Nov 2020.
19. Federhen S. The NCBI taxonomy database. *Nucleic Acids Res*. 2012; 40(Database issue):D136–43.
20. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, et al. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res*. 2015; 43(Database issue):D36–42.
21. UniProt Consortium T, Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2017;45:D158–69.
22. wwPDB consortium, Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, et al. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res*. 2019;47:D520–8.
23. Waagmeester A, Kutmon M, Riutta A, Miller R, Willighagen EL, Evelo CT, et al. Using the semantic web for rapid integration of WikiPathways with other biological online data resources. *PLoS Comput Biol*. 2016;12: e1004989.
24. Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2020;48:D9–16.
25. Thornton K, Solbrig H, Stupp GS, Labra Gayo JE, Mietchen D, Prud'hommeaux E, et al. Using Shape Expressions (ShEx) to share RDF data models and to guide curation with rigorous validation. In: Hitzler P, Fernández M, Janowicz K, Zaveri A, Gray AJG, Lopez V, et al., editors. *The Semantic Web*. Cham: Springer International Publishing; 2019. p. 606–620. doi:[https://doi.org/10.1007/978-3-030-21348-0\\_39](https://doi.org/10.1007/978-3-030-21348-0_39).
26. Prud'hommeaux E, Labra Gayo JE, Solbrig H. Shape expressions: an RDF validation and transformation language. In: *Proceedings of the 10th international conference on semantic systems*. New York: Association for Computing Machinery; 2014. p. 32–40. <https://doi.org/10.1145/2660517.2660523>.
27. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*. 2020;382: 727–33.
28. Berners-Lee T, Hendler J, Lassila O. The semantic web. *Sci Am*. 2001;284: 34–43.
29. SPARQL 1.1 Query Language. <https://www.w3.org/TR/sparql11-query/>. Accessed 25 Nov 2020.
30. OWL 2 Web Ontology Language Document Overview (Second Edition). <https://www.w3.org/TR/owl2-overview/>. Accessed 25 Nov 2020.
31. Linked Data - Design Issues. <https://www.w3.org/DesignIssues/LinkedData.html>. Accessed 25 Nov 2020.
32. The Linked Open Data Cloud. <https://lod-cloud.net/>. Accessed 25 Nov 2020.
33. Samwald M, Jentzsch A, Bouton C, Kallesøe CS, Willighagen E, Hajagos J, et al. Linked open drug data for pharmaceutical research and development. *J Cheminformatics*. 2011;3:19.
34. Help:Statements - Wikidata. <https://www.wikidata.org/wiki/Help:Statements>. Accessed 25 Nov 2020.
35. Hernandez D, Hogan A, Kroetzsch M. Reifying RDF: what works well with Wikidata? p. 16.
36. virus taxon (E192) - Wikidata. <https://www.wikidata.org/wiki/EntitySchema:E192>. Accessed 30 Nov 2020.
37. strain (E174) - Wikidata. <https://www.wikidata.org/wiki/EntitySchema:E174>. Accessed 27 Nov 2020.
38. disease (E69) - Wikidata. <https://www.wikidata.org/wiki/EntitySchema:E69>. Accessed 27 Nov 2020.
39. virus strain (E170) - Wikidata. <https://www.wikidata.org/wiki/EntitySchema:E170>. Accessed 27 Nov 2020.
40. virus gene (E165) - Wikidata. <https://www.wikidata.org/wiki/EntitySchema:E165>. Accessed 27 Nov 2020.

