Wikidata and COVID-19

Creating a collaborative knowledge graph from CORD-19 scholarly publications

Houcemeddine Turki, *University of Sfax, Tunisia*

جامعة صفاقس
University of Sfax

doi:10.5281/zenodo.4048515

W3C

Ai2 Allen Institute for AI

Team Members
WikiProject COVID-19

Houcemeddine Turki, *University of Sfax, Tunisia*
Thomas Shafee, *La Trobe University, Australia*
Daniel Mietchen, *University of Virginia, United States of America*
Tiago Lubiana, *University of São Paolo, Brazil*
Dariusz Jemielniak, *Kozminski University, Poland*
Jose Emilio Labra Gayo, *University of Oviedo, Spain*
Eric Prud'Hommeaux, *World Wide Web Consortium, United States of America*
Mohamed Ali Hadj Taieb, *University of Sfax, Tunisia*
Mohamed Ben Aouicha, *University of Sfax, Tunisia*
Mus'ab Banat, *Hashemite University, Jordan*
Diptanshu Das, *Institute of Child Health, India*

جامعة صفاقس
University of Sfax

University of Sfax

- Top 2 in Tunisia
- Ranked among the first African universities in Computer Science (1st in Leiden CWTS, 2nd in URAP)
- Research Group specialized in Semantic Technologies and Biomedical Data Science (Data Engineering and Semantics)
- Technopark dealing with computer science research

About Us

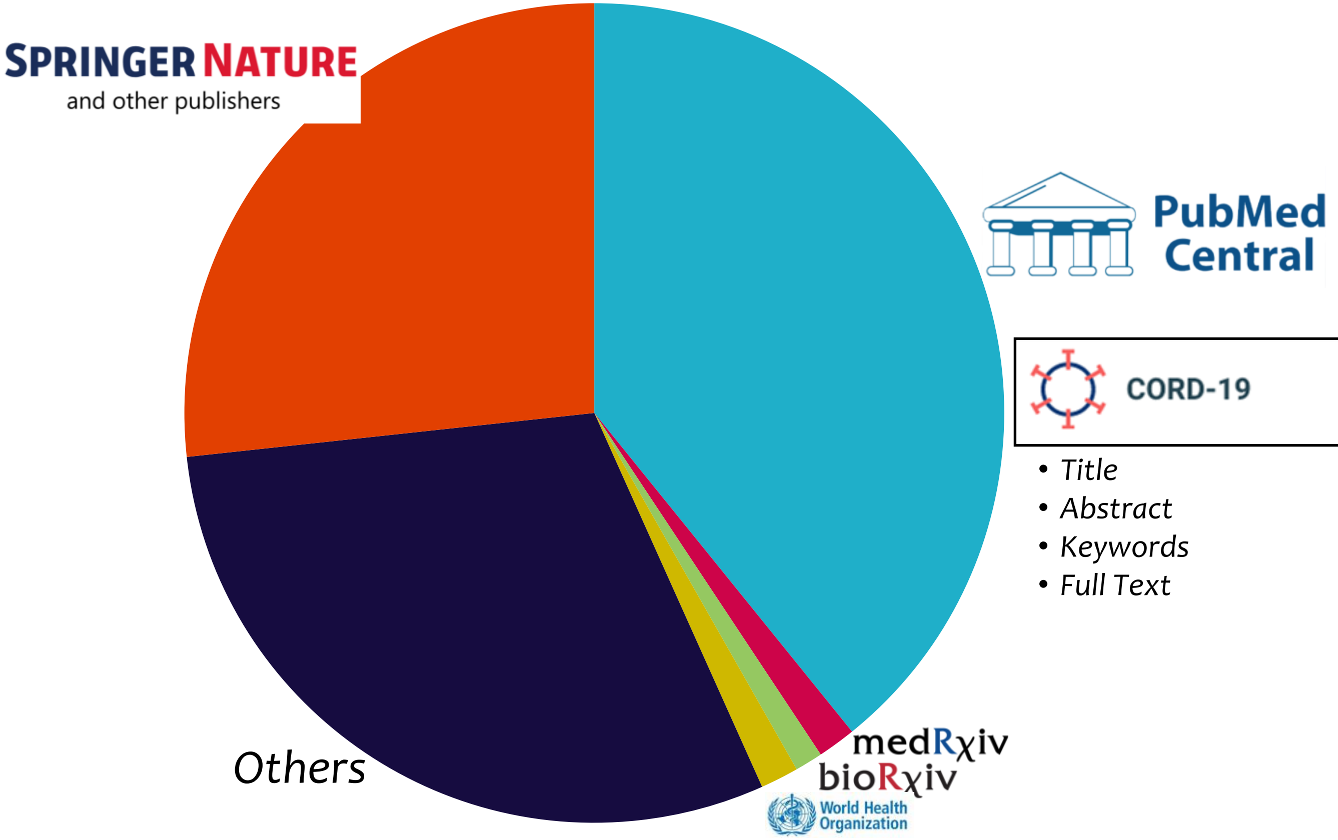
WikiProject COVID-19
University of Sfax

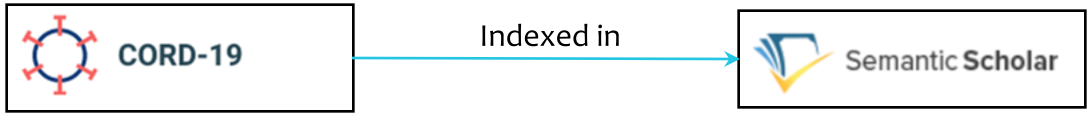
Introduction

Insights on Covid-19 Open Research Dataset (CORD-19)

CORD-19: COVID-19 Open Research Dataset

An initiative of *AI2*, *CZI*, *MSR*, *Georgetown*, *NIH*, and *The White House*





- *Title*
- *Abstract*
- *Keywords*
- *Full Text*

CORD-19

The sum of all human knowledge about COVID-19

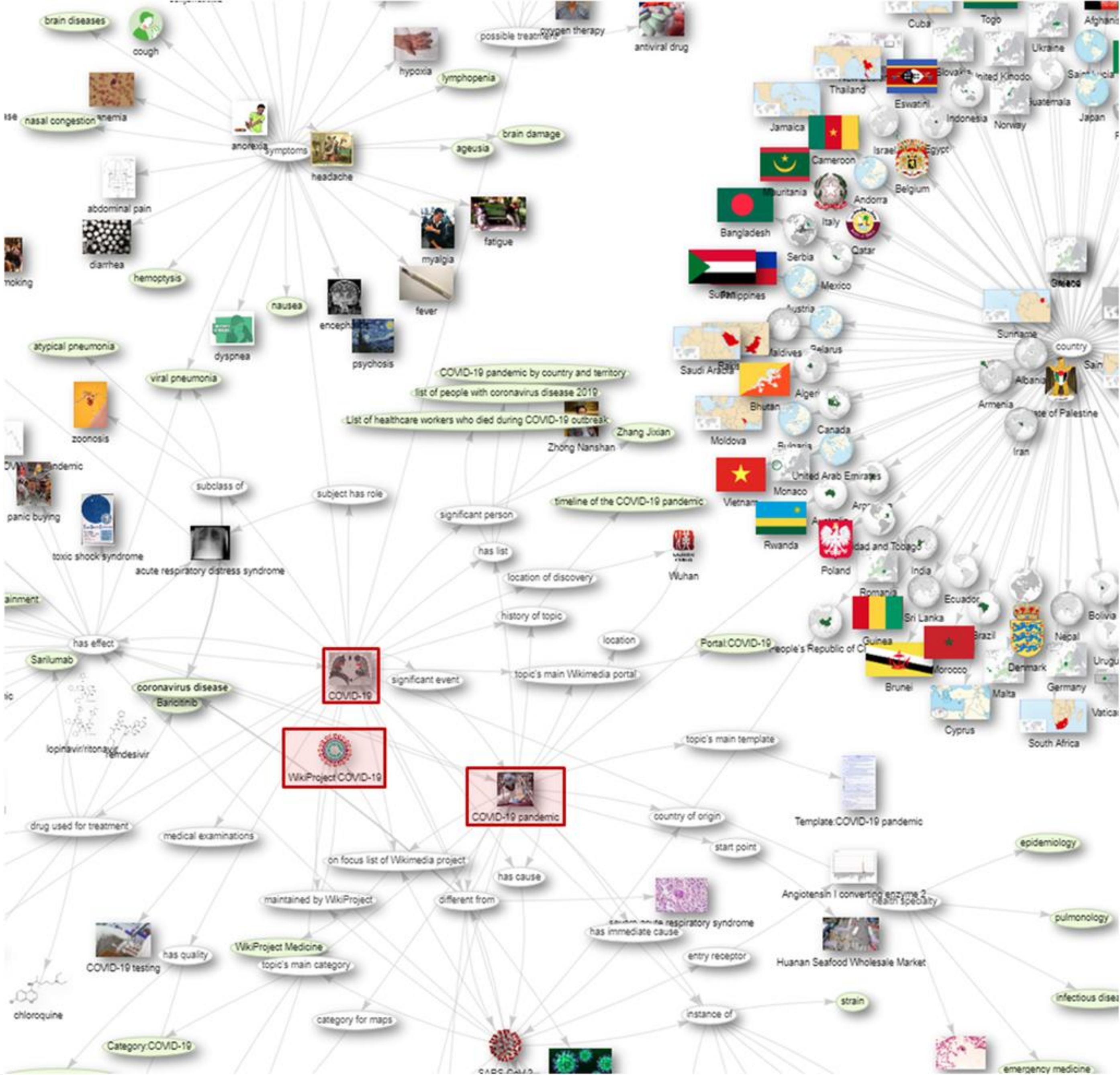
A huge amount of raw texts

- Difficult to study by humans
- Hard to process by computer programs
- Knowledge updated every day

Proposed Solution

A knowledge graph for COVID-19 information

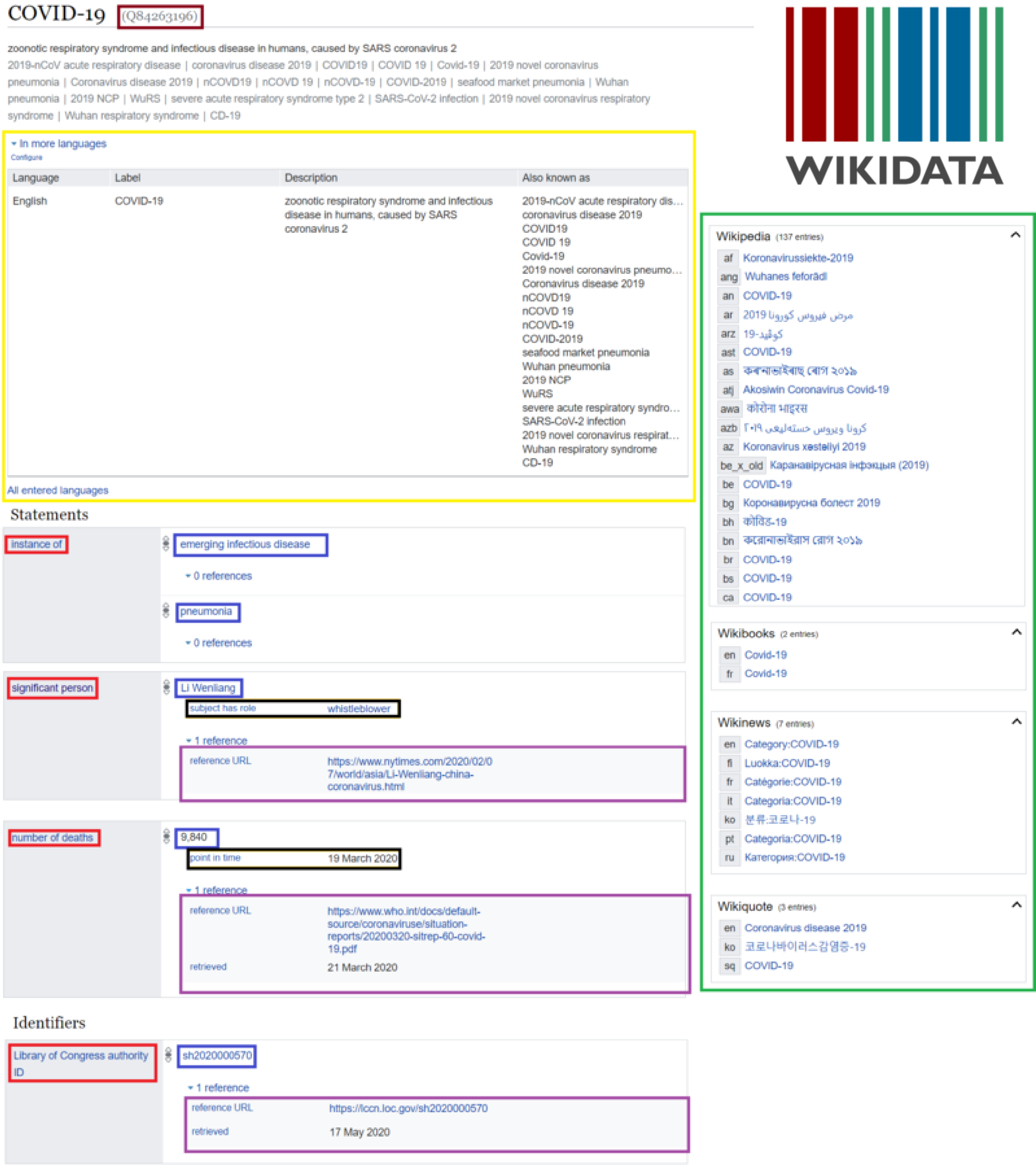
- A fully structured semantic database in the form of RDF triples.
- Human-readable, Machine-readable
- Findable, Accessible, Interoperable and Reusable
- Flexible data model for the representation of COVID-19 information
- Screened using SPARQL



Wikidata

A large-scale free knowledge base

- Available at https://www.wikidata.org
- Items and properties are assigned language-independent identifiers and labels and descriptions in multiple languages
- Items and properties are assigned statements in the form of RDF triples. These statements can be detailed using triple qualifiers and references.
- Statements can be relational (taxonomic or non-taxonomic) or non-relational ones (objects as values, external IDs, URLs and dates…).
- CC0 License (easily reusable but cannot include various datasets released under CC-BY and other licenses).



COVID-19 (Q84263196)

zoonotic respiratory syndrome and infectious disease in humans, caused by SARS coronavirus 2

2019-nCoV acute respiratory disease | coronavirus disease 2019 | COVID19 | COVID 19 | Covid-19 | 2019 novel coronavirus pneumonia | Coronavirus disease 2019 | nCOVD19 | nCOVD 19 | nCOVD-19 | COVID-2019 | seafood market pneumonia | Wuhan pneumonia | 2019 NCP | WuRS | severe acute respiratory syndrome type 2 | SARS-CoV-2 infection | 2019 novel coronavirus respiratory syndrome | Wuhan respiratory syndrome | CD-19

▾ In more languages
Configure

Language	Label	Description	Also known as
English	COVID-19	zoonotic respiratory syndrome and infectious disease in humans, caused by SARS coronavirus 2	2019-nCoV acute respiratory dis… coronavirus disease 2019 COVID19 COVID 19 Covid-19 2019 novel coronavirus pneumo… Coronavirus disease 2019 nCOVD19 nCOVD 19 nCOVD-19 COVID-2019 seafood market pneumonia Wuhan pneumonia 2019 NCP WuRS severe acute respiratory syndro… SARS-CoV-2 infection 2019 novel coronavirus respirat… Wuhan respiratory syndrome CD-19

All entered languages

Statements

instance of
- emerging infectious disease
  - ▾ 0 references
- pneumonia
  - ▾ 0 references

significant person
- Li Wenliang
  - subject has role: whistleblower
  - ▾ 1 reference
    - reference URL: https://www.nytimes.com/2020/02/07/world/asia/Li-Wenliang-china-coronavirus.html

number of deaths
- 9,840
  - point in time: 19 March 2020
  - ▾ 1 reference
    - reference URL: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200320-sitrep-60-covid-19.pdf
    - retrieved: 21 March 2020

Identifiers

Library of Congress authority ID
- sh2020000570
  - ▾ 1 reference
    - reference URL: https://lccn.loc.gov/sh2020000570
    - retrieved: 17 May 2020

Wikipedia (137 entries)

af	Koronavirussiekte-2019
ang	Wuhanes feforādl
an	COVID-19
ar	مرض فيروس كورونا 2019
arz	كوفيد-19
ast	COVID-19
as	কৰ'নাভাইৰাছ ৰোগ ২০১৯
atj	Akosiwin Coronavirus Covid-19
awa	कोरोना भाइरस
azb	کرونا ویروس خسته‌لیغی ۲۰۱۹
az	Koronavirus xəstəliyi 2019
be_x_old	Каранавірусная інфэкцыя (2019)
be	COVID-19
bg	Коронавирусна болест 2019
bh	कोविड-19
bn	করোনাভাইরাস রোগ ২০১৯
br	COVID-19
bs	COVID-19
ca	COVID-19

Wikibooks (2 entries)

en	Covid-19
fr	Covid-19

Wikinews (7 entries)

en	Category:COVID-19
fi	Luokka:COVID-19
fr	Catégorie:COVID-19
it	Categoria:COVID-19
ko	분류:코로나-19
pt	Categoria:COVID-19
ru	Категория:COVID-19

Wikiquote (3 entries)

en	Coronavirus disease 2019
ko	코로나바이러스감염증-19
sq	COVID-19

WIKIDATA

Comparison

Collaborative multidisciplinary knowledge graphs vs. Specialized knowledge graphs

Collaborative multidisciplinary knowledge graphs
Wikidata

- Include already available non-COVID-19 items and statements
  - Allow correlation analysis between COVID-19 information and non-COVID-19 information
  - Data models already existing in part
- Edited and curated by a large community of editors

Specialized knowledge graphs
CORD-19 NEKG

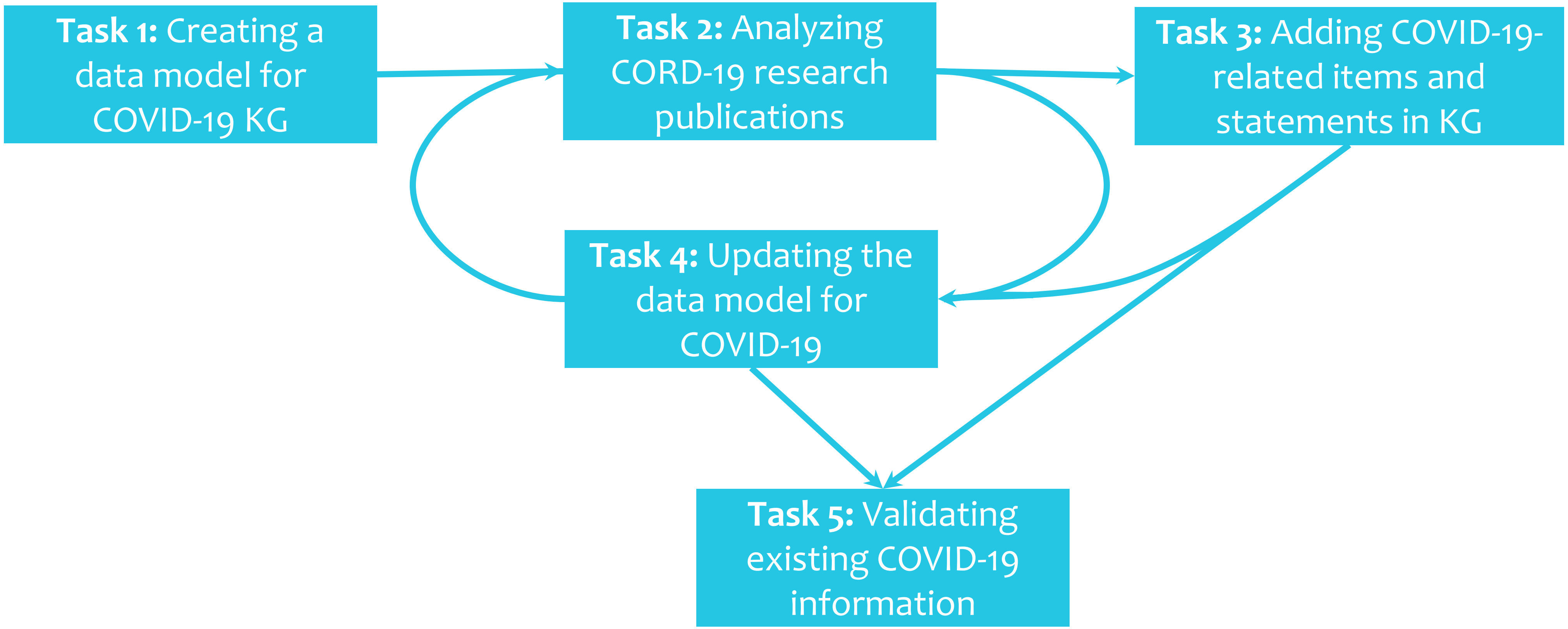
- Does not include already available non-COVID-19 items and statements
  - Only deep analysis of COVID-19 information
  - Data models developed from scratch
- Edited and curated by a panel of experts

Methods

Creating a large-scale COVID-19 knowledge graph in Wikidata

Roadmap

Creation of knowledge graphs from CORD-19



Task 1: Creating a data model for COVID-19 KG

Task 2: Analyzing CORD-19 research publications

Task 3: Adding COVID-19-related items and statements in KG

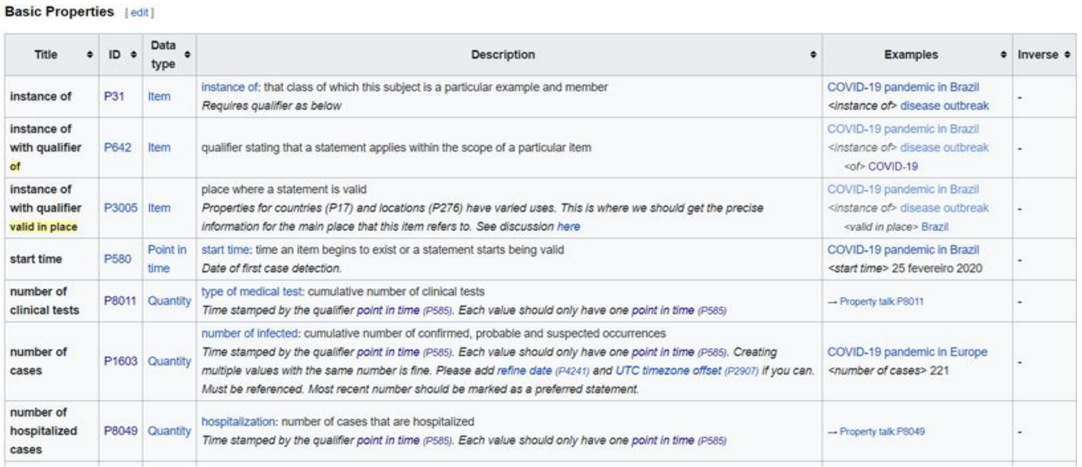
Task 4: Updating the data model for COVID-19

Task 5: Validating existing COVID-19 information

Creating a data model for COVID-19 KG

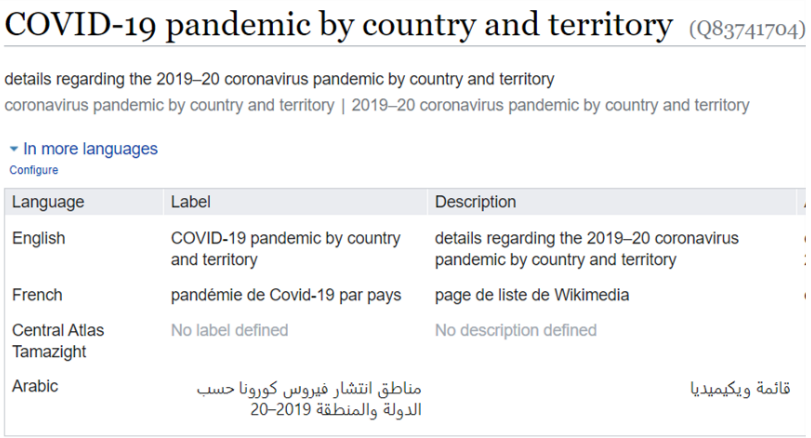
Task 1

- Defining the classes for COVID-19-related items (https://www.wikidata.org/wiki/Special:NewItem)
- Defining the structure of the items of each class using ShEx (https://www.wikidata.org/wiki/Wikidata:WikiProject_COVID-19/Data_models)
- Adding and sustaining Wikidata properties to characterize COVID-19 statements (https://www.wikidata.org/wiki/Wikidata:Property_proposal)
- CORD-19 NEKG: https://github.com/Wimmics/CovidOnTheWeb/blob/master/doc/01-data-modeling.md (YouTube: https://youtu.be/oUk9PXGM2fY)



Basic Properties [edit]

Title	ID	Data type	Description	Examples	Inverse
instance of	P31	Item	instance of: that class of which this subject is a particular example and member *Requires qualifier as below*	COVID-19 pandemic in Brazil *<instance of>* disease outbreak	-
instance of with qualifier of	P642	Item	qualifier stating that a statement applies within the scope of a particular item	COVID-19 pandemic in Brazil *<instance of>* disease outbreak *<of>* COVID-19	-
instance of with qualifier valid in place	P3005	Item	place where a statement is valid *Properties for countries (P17) and locations (P276) have varied uses. This is where we should get the precise information for the main place that this item refers to. See discussion here*	COVID-19 pandemic in Brazil *<instance of>* disease outbreak *<valid in place>* Brazil	-
start time	P580	Point in time	start time: time an item begins to exist or a statement starts being valid *Date of first case detection.*	COVID-19 pandemic in Brazil *<start time>* 25 fevereiro 2020	-
number of clinical tests	P8011	Quantity	type of medical test: cumulative number of clinical tests *Time stamped by the qualifier point in time (P585). Each value should only have one point in time (P585)*	→ Property talk:P8011	-
number of cases	P1603	Quantity	number of infected: cumulative number of confirmed, probable and suspected occurrences *Time stamped by the qualifier point in time (P585). Each value should only have one point in time (P585). Creating multiple values with the same number is fine. Please add refine date (P4241) and UTC timezone offset (P2907) if you can. Must be referenced. Most recent number should be marked as a preferred statement.*	COVID-19 pandemic in Europe *<number of cases>* 221	-
number of hospitalized cases	P8049	Quantity	hospitalization: number of cases that are hospitalized *Time stamped by the qualifier point in time (P585). Each value should only have one point in time (P585)*	→ Property talk:P8049	-



COVID-19 pandemic by country and territory (Q83741704)

details regarding the 2019–20 coronavirus pandemic by country and territory

coronavirus pandemic by country and territory | 2019–20 coronavirus pandemic by country and territory

▾ In more languages

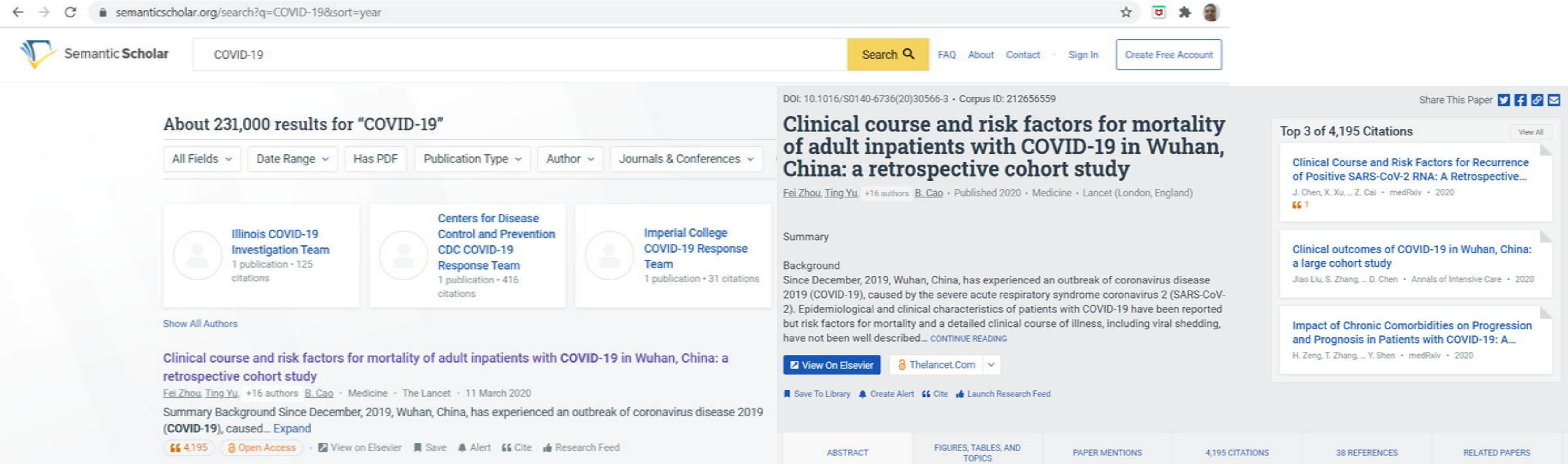
Configure

Language	Label	Description
English	COVID-19 pandemic by country and territory	details regarding the 2019–20 coronavirus pandemic by country and territory
French	pandémie de Covid-19 par pays	page de liste de Wikimedia
Central Atlas Tamazight	No label defined	No description defined
Arabic	مناطق انتشار فيروس كورونا حسب الدولة والمنطقة 2019–20	قائمة ويكيميديا

Analyzing CORD-19 research publications

Task 2

- Human screening of CORD-19 research publications on Semantic Scholar and PubMed Central for the creation and adjustment of COVID-19-related statements on Wikidata.



About 231,000 results for "COVID-19"

Illinois COVID-19 Investigation Team
1 publication • 125 citations

Centers for Disease Control and Prevention CDC COVID-19 Response Team
1 publication • 416 citations

Imperial College COVID-19 Response Team
1 publication • 31 citations

Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study

Fei Zhou, Ting Yu, +16 authors B. Cao · Medicine · The Lancet · 11 March 2020

Summary Background Since December, 2019, Wuhan, China, has experienced an outbreak of coronavirus disease 2019 (**COVID-19**), caused… Expand

4,195 · Open Access · View on Elsevier · Save · Alert · Cite · Research Feed

DOI: 10.1016/S0140-6736(20)30566-3 • Corpus ID: 212656559

Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study

Fei Zhou, Ting Yu, +16 authors B. Cao • Published 2020 • Medicine • Lancet (London, England)

Summary

Background

Since December, 2019, Wuhan, China, has experienced an outbreak of coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Epidemiological and clinical characteristics of patients with COVID-19 have been reported but risk factors for mortality and a detailed clinical course of illness, including viral shedding, have not been well described… CONTINUE READING

Top 3 of 4,195 Citations

Clinical Course and Risk Factors for Recurrence of Positive SARS-CoV-2 RNA: A Retrospective…
J. Chen, X. Xu, … Z. Cai • medRxiv • 2020
1

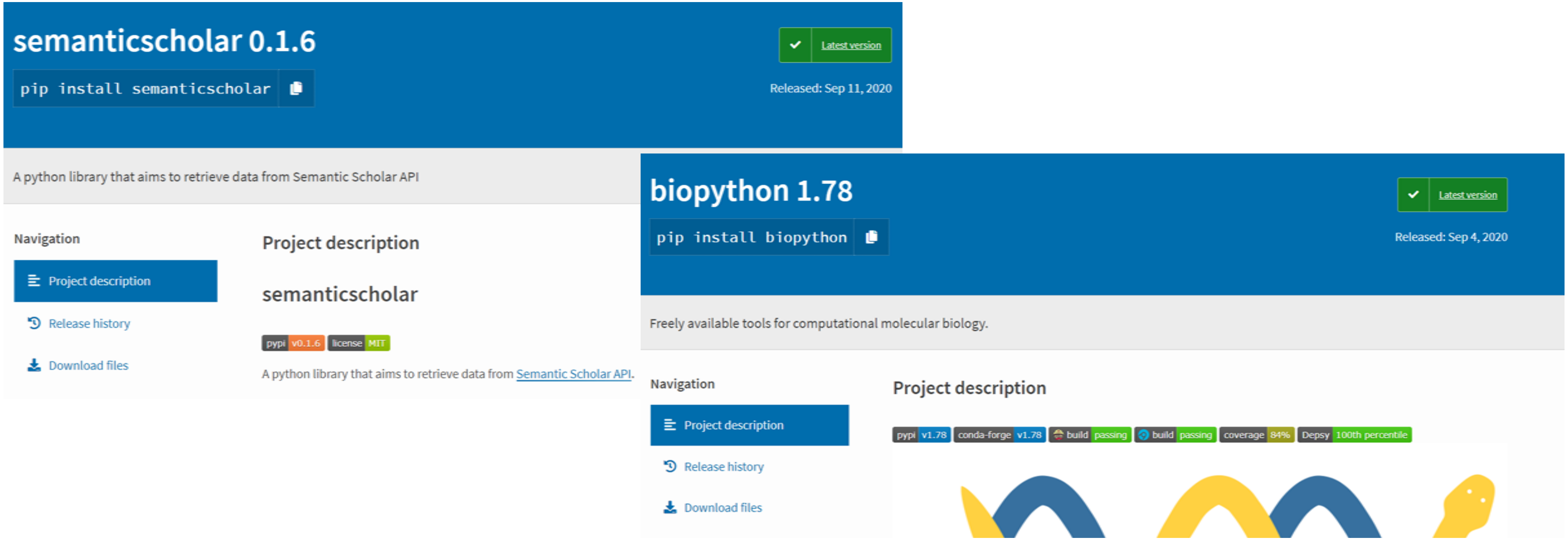
Clinical outcomes of COVID-19 in Wuhan, China: a large cohort study
Jiao Liu, S. Zhang, … D. Chen • Annals of Intensive Care • 2020

Impact of Chronic Comorbidities on Progression and Prognosis in Patients with COVID-19: A…
H. Zeng, T. Zhang, … Y. Shen • medRxiv • 2020

Analyzing CORD-19 research publications

Task 2

- Processing CORD-19 research publications using Semantic Scholar API and PubMed Central Entrez API and corresponding Python Libraries.



semanticscholar 0.1.6

```
pip install semanticscholar
```

Released: Sep 11, 2020

A python library that aims to retrieve data from Semantic Scholar API

Project description

semanticscholar

A python library that aims to retrieve data from Semantic Scholar API.

biopython 1.78

```
pip install biopython
```

Released: Sep 4, 2020

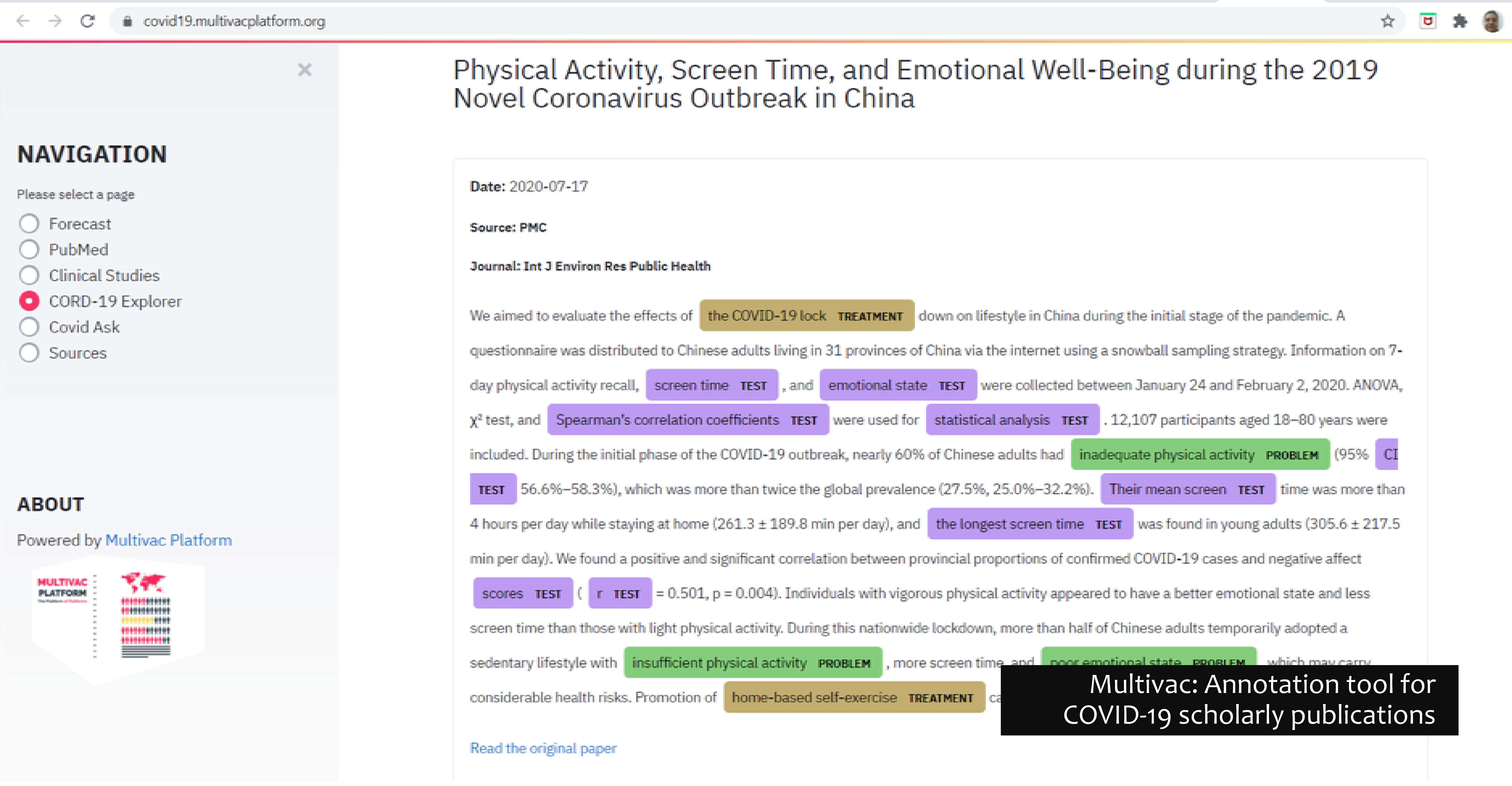
Freely available tools for computational molecular biology.

Project description

Analyzing CORD-19 research publications

Task 2

- Annotating CORD-19 research publications with named entities from semantic databases such as Wikidata.
  - Eliminating stop words
  - Extracting n-grams
  - Finding n-grams in knowledge graphs using APIs
- The annotation process can be contextualized and restricted to the items included in a given class of a knowledge graph (Diseases, drugs, etc…).



NAVIGATION

Please select a page

○ Forecast
○ PubMed
○ Clinical Studies
● CORD-19 Explorer
○ Covid Ask
○ Sources

ABOUT

Powered by Multivac Platform

Physical Activity, Screen Time, and Emotional Well-Being during the 2019 Novel Coronavirus Outbreak in China

Date: 2020-07-17

Source: PMC

Journal: Int J Environ Res Public Health

We aimed to evaluate the effects of the COVID-19 lock TREATMENT down on lifestyle in China during the initial stage of the pandemic. A questionnaire was distributed to Chinese adults living in 31 provinces of China via the internet using a snowball sampling strategy. Information on 7-day physical activity recall, screen time TEST , and emotional state TEST were collected between January 24 and February 2, 2020. ANOVA, χ² test, and Spearman's correlation coefficients TEST were used for statistical analysis TEST . 12,107 participants aged 18–80 years were included. During the initial phase of the COVID-19 outbreak, nearly 60% of Chinese adults had inadequate physical activity PROBLEM (95% CI TEST 56.6%–58.3%), which was more than twice the global prevalence (27.5%, 25.0%–32.2%). Their mean screen TEST time was more than 4 hours per day while staying at home (261.3 ± 189.8 min per day), and the longest screen time TEST was found in young adults (305.6 ± 217.5 min per day). We found a positive and significant correlation between provincial proportions of confirmed COVID-19 cases and negative affect scores TEST (r TEST = 0.501, p = 0.004). Individuals with vigorous physical activity appeared to have a better emotional state and less screen time than those with light physical activity. During this nationwide lockdown, more than half of Chinese adults temporarily adopted a sedentary lifestyle with insufficient physical activity PROBLEM , more screen time, and poor emotional state PROBLEM which may carry considerable health risks. Promotion of home-based self-exercise TREATMENT ca

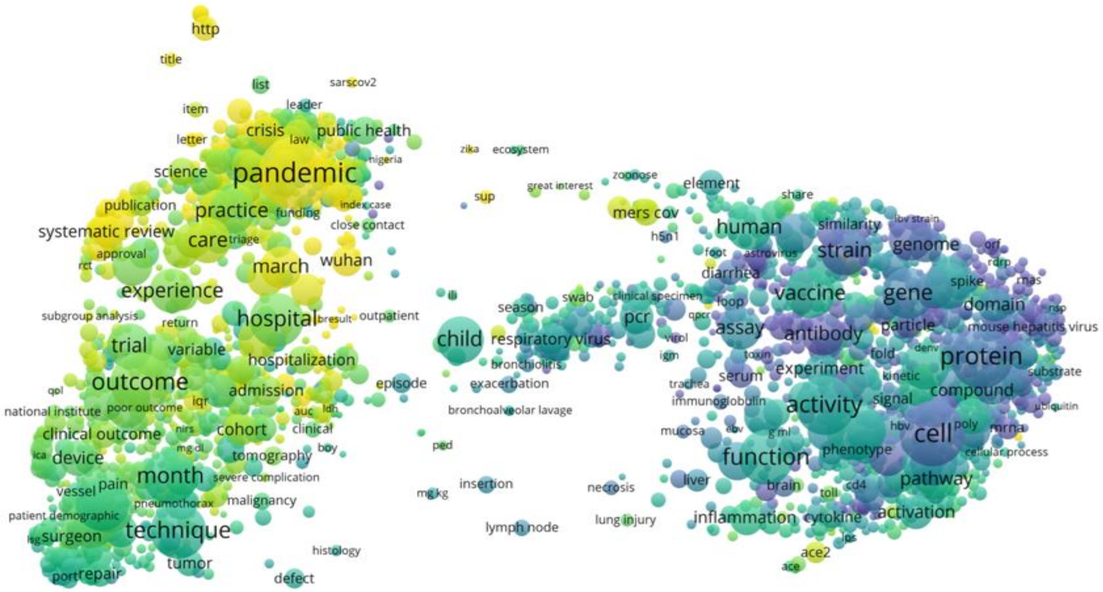
Read the original paper

Multivac: Annotation tool for COVID-19 scholarly publications

Analyzing CORD-19 research publications

Task 2

- Using Topic Modelling (Latent Dirichlet Allocation) of CORD-19 scholarly publications to retrieve the main concepts related to COVID-19 pandemic.
- Redundancy of returned concepts can be solved using word similarity metrics such as semantic similarity measures and word embeddings.



Analyzing CORD-19 research publications

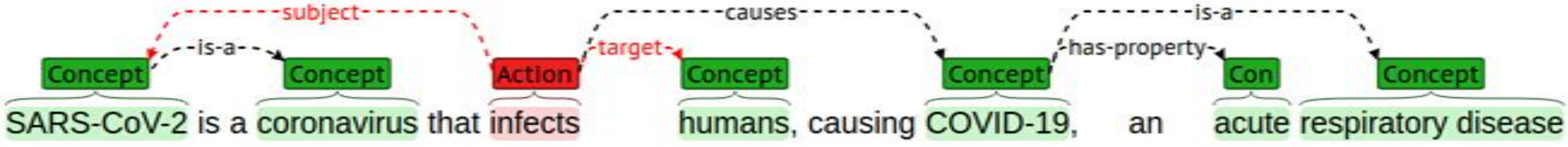
Task 2

- Annotating CORD-19 research publications with relations from semantic databases such as Wikidata.
- Various methods:
  - Human screening and annotation of CORD-19 using a tool such as https://brat.nlplab.org/.
  - Using a benchmark of semantic relations of the same types (particularly drug interactions and drug-disease relations) with ML techniques (CNN, RNN and LSTM) or word embeddings (Word2Vec, BERT and ELMo) to annotate and retrieve semantic relations from CORD-19.
  - Inferring semantic relations for topic modelling of CORD-19.

Analyzing CORD-19 research publications

Task 2

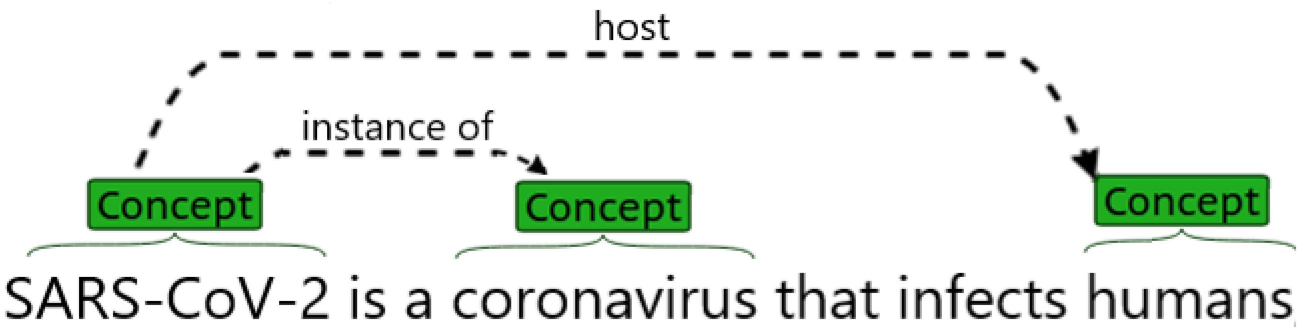
- Two data models of relation annotation:
  - SAT+R: *Subject*, *Action*, *Target* triplets and additional *Relations*
    - Limited number of relation types linking annotated entities
    - Supported relations are generic (Subject, Target, Has property…)
    - Semantic links between concepts (*Green*) are extracted from analyzed text and annotated as action entities (*Red*).


subject
is-a
causes
target
is-a
has-property
Concept
Concept
Action
Concept
Concept
Con
Concept
SARS-CoV-2 is a coronavirus that infects humans, causing COVID-19, an acute respiratory disease

Analyzing CORD-19 research publications

Task 2

- Two data models of relation annotation:
  - Subject-Object relation annotations
    - Every type of biomedical information is represented by a relation type
    - Only identified concepts are annotated in the analyzed text
    - Actions are represented as links between concepts and are not consequently annotated as entities.



Analyzing CORD-19 research publications

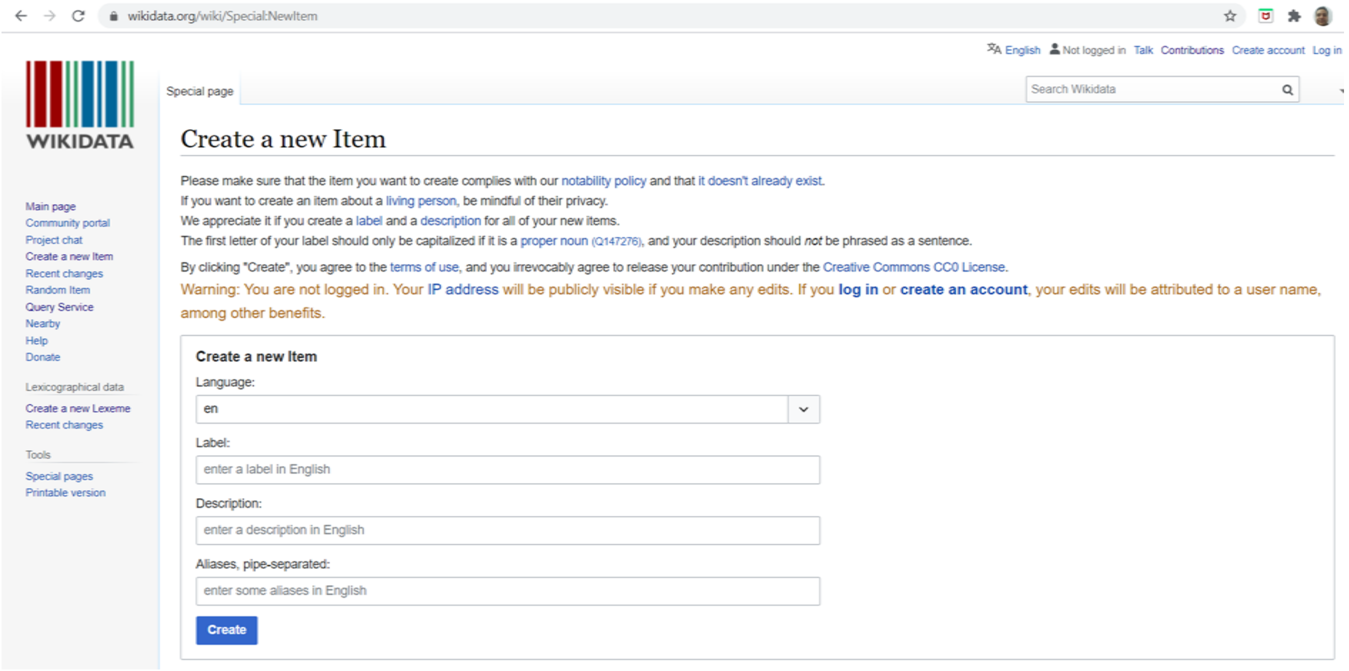
Task 2

- Using Bibliometric-Enhanced Information Retrieval to enhance the extraction of biomedical relations
  - MeSH Keywords
    - PubMed Records include MeSH Keywords describing the output of the corresponding research publication. MeSH Keywords involve a Heading describing a biomedical concept and a qualifier specifying the studied pattern in the Heading/Qualifier form.
    - A combination of two MeSH Keywords can be used to infer a given biomedical relation.
    - Example: "Hepatitis C/therapy" and "Sofosbuvir/therapeutic use“ in https://pubmed.ncbi.nlm.nih.gov/32526210/ can be used to determine that Sofosbuvir is used as a drug for Hepatitis C.
  - Publication types
    - Section titles in literature reviews can provide an idea about the types of available relations in each section of the reviews.
    - Example: “Drug interactions” section in a literature review about “Azithromycin” involves semantic relations about significant drug interactions of Azithromycin.

Adding COVID-19-related items and statements in KG

Task 3

- Human creation of COVID-19-related items and statements
  - Item creation: https://www.wikidata.org/wiki/Special:NewItem
  - Statement creation: https://www.wikidata.org/wiki/<item>



Create a new Item

Please make sure that the item you want to create complies with our notability policy and that it doesn't already exist.
If you want to create an item about a living person, be mindful of their privacy.
We appreciate it if you create a label and a description for all of your new items.
The first letter of your label should only be capitalized if it is a proper noun (Q147276), and your description should *not* be phrased as a sentence.

By clicking "Create", you agree to the terms of use, and you irrevocably agree to release your contribution under the Creative Commons CC0 License.

Warning: You are not logged in. Your IP address will be publicly visible if you make any edits. If you **log in** or **create an account**, your edits will be attributed to a user name, among other benefits.

Create a new Item

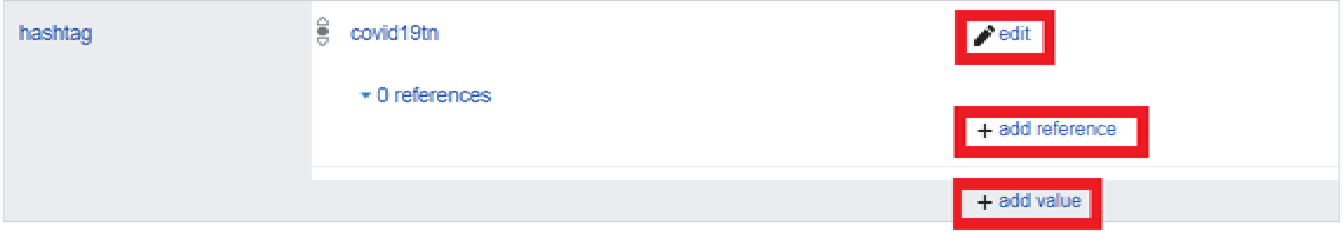
Language: en

Label: enter a label in English

Description: enter a description in English

Aliases, pipe-separated: enter some aliases in English

Create

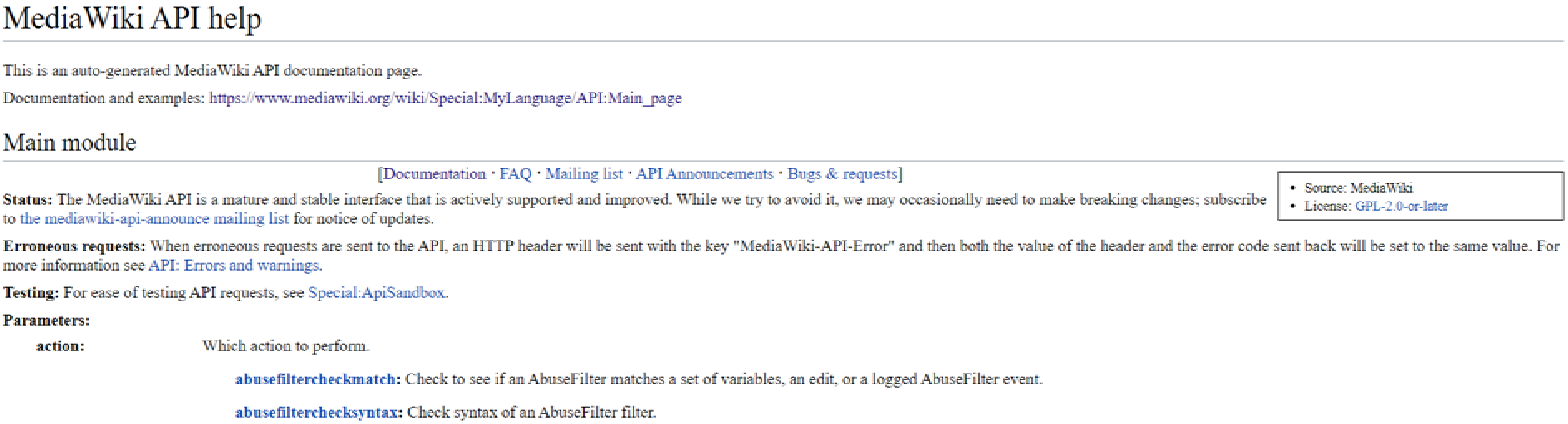


hashtag: covid19tn

Adding COVID-19-related items and statements in KG

Task 3

- Wikidata API (https://www.wikidata.org/w/api.php)
  - Mass extraction of COVID-19 multidisciplinary information
  - Mass adjustment and modification of COVID-19 multidisciplinary information



MediaWiki API help

This is an auto-generated MediaWiki API documentation page.

Documentation and examples: https://www.mediawiki.org/wiki/Special:MyLanguage/API:Main_page

Main module

[Documentation · FAQ · Mailing list · API Announcements · Bugs & requests]

- Source: MediaWiki
- License: GPL-2.0-or-later

Status: The MediaWiki API is a mature and stable interface that is actively supported and improved. While we try to avoid it, we may occasionally need to make breaking changes; subscribe to the mediawiki-api-announce mailing list for notice of updates.

Erroneous requests: When erroneous requests are sent to the API, an HTTP header will be sent with the key "MediaWiki-API-Error" and then both the value of the header and the error code sent back will be set to the same value. For more information see API: Errors and warnings.

Testing: For ease of testing API requests, see Special:ApiSandbox.

Parameters:

action: Which action to perform.

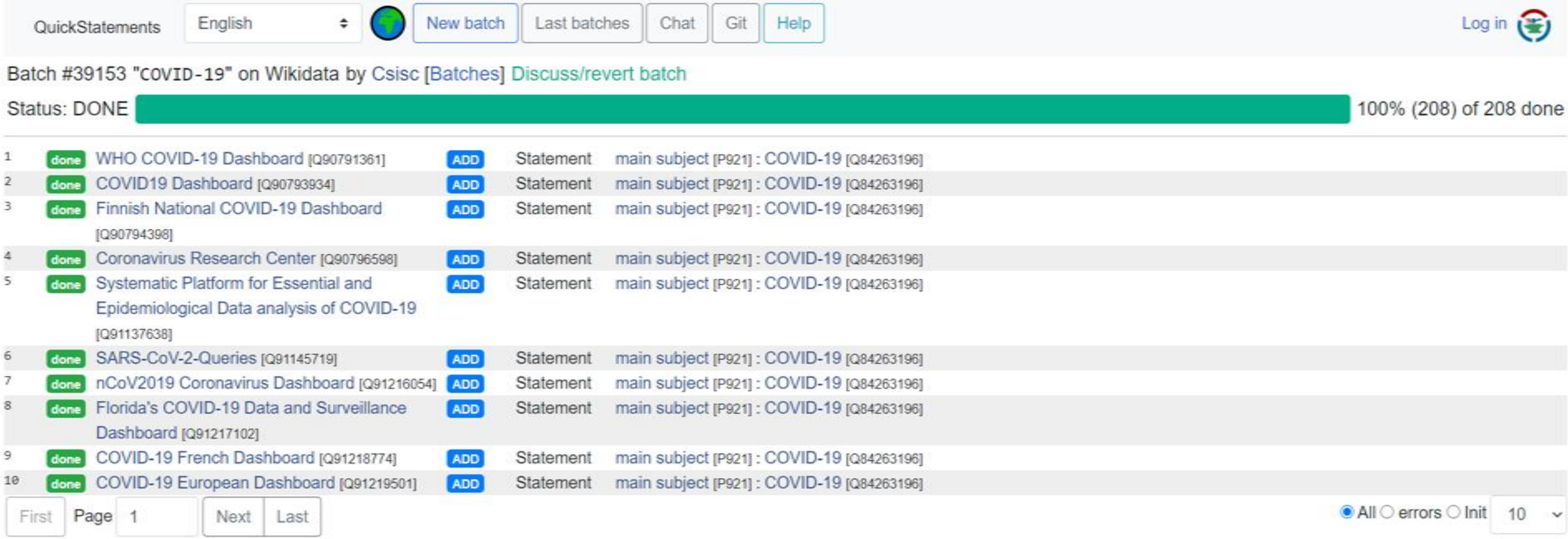
abusefiltercheckmatch: Check to see if an AbuseFilter matches a set of variables, an edit, or a logged AbuseFilter event.

abusefilterchecksyntax: Check syntax of an AbuseFilter filter.

Adding COVID-19-related items and statements in KG

Task 3

- QuickStatements (https://quickstatements.toolforge.org)
  - Upload batches of semantic relations to Wikidata
  - Can be programmatically used



Batch #39153 "COVID-19" on Wikidata by Csisc [Batches] Discuss/revert batch

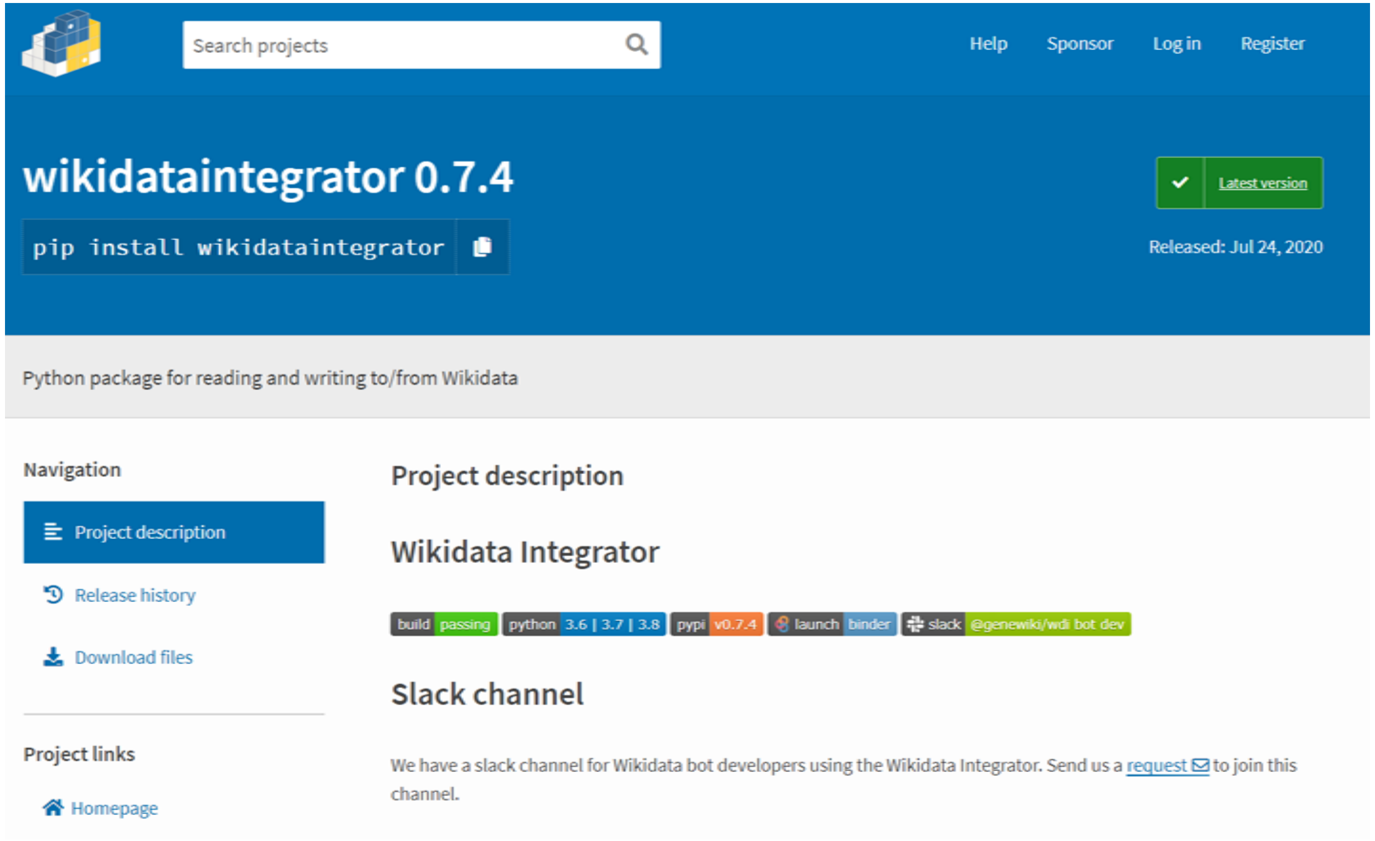
Status: DONE 100% (208) of 208 done

1	done	WHO COVID-19 Dashboard [Q90791361]	ADD	Statement	main subject [P921] : COVID-19 [Q84263196]
2	done	COVID19 Dashboard [Q90793934]	ADD	Statement	main subject [P921] : COVID-19 [Q84263196]
3	done	Finnish National COVID-19 Dashboard [Q90794398]	ADD	Statement	main subject [P921] : COVID-19 [Q84263196]
4	done	Coronavirus Research Center [Q90796598]	ADD	Statement	main subject [P921] : COVID-19 [Q84263196]
5	done	Systematic Platform for Essential and Epidemiological Data analysis of COVID-19 [Q91137638]	ADD	Statement	main subject [P921] : COVID-19 [Q84263196]
6	done	SARS-CoV-2-Queries [Q91145719]	ADD	Statement	main subject [P921] : COVID-19 [Q84263196]
7	done	nCoV2019 Coronavirus Dashboard [Q91216054]	ADD	Statement	main subject [P921] : COVID-19 [Q84263196]
8	done	Florida's COVID-19 Data and Surveillance Dashboard [Q91217102]	ADD	Statement	main subject [P921] : COVID-19 [Q84263196]
9	done	COVID-19 French Dashboard [Q91218774]	ADD	Statement	main subject [P921] : COVID-19 [Q84263196]
10	done	COVID-19 European Dashboard [Q91219501]	ADD	Statement	main subject [P921] : COVID-19 [Q84263196]

Adding COVID-19-related items and statements in KG

Task 3

- Wikidata Integrator (https://pypi.org/project/wikidataintegrator/)
  - A Python Library to analyze, add and adjust Wikidata statements



wikidataintegrator 0.7.4

```
pip install wikidataintegrator
```

Released: Jul 24, 2020

Python package for reading and writing to/from Wikidata

Project description

Wikidata Integrator

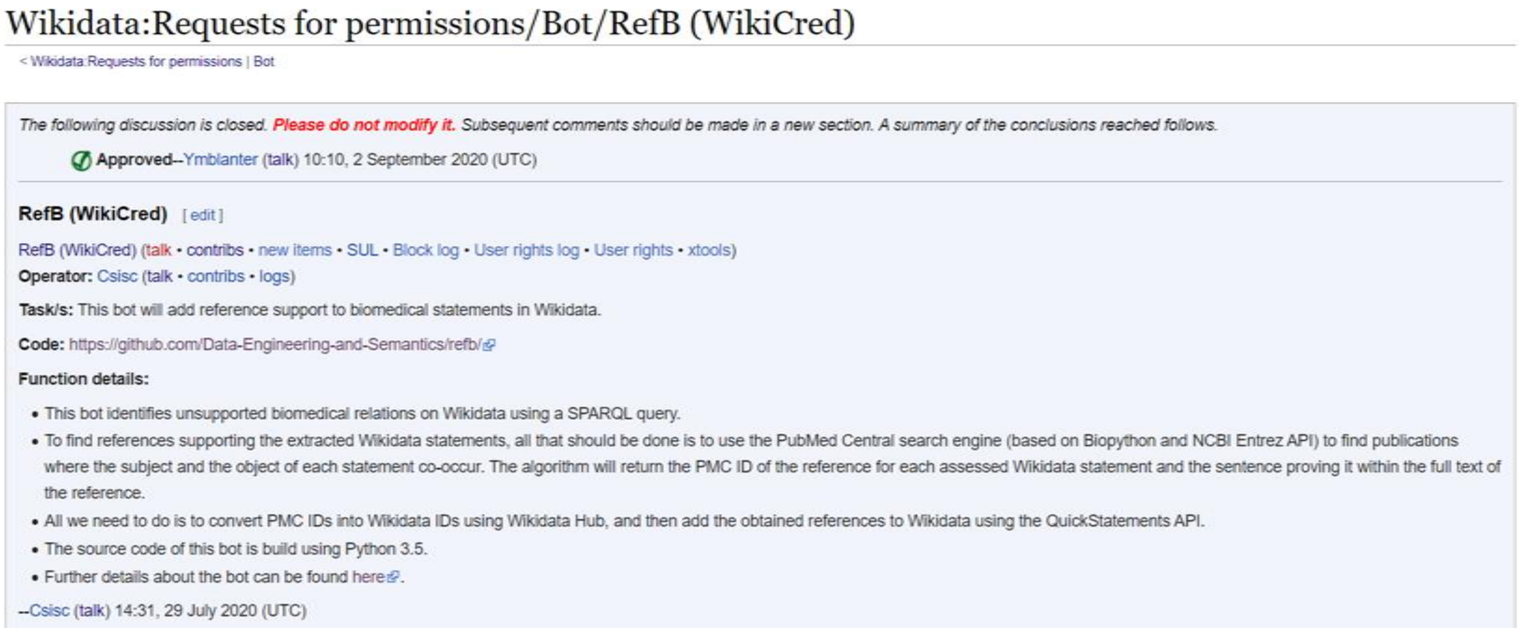
Slack channel

We have a slack channel for Wikidata bot developers using the Wikidata Integrator. Send us a request to join this channel.

Adding COVID-19-related items and statements in KG

Task 3

- To apply a bot on Wikidata
  - Create a Python code
  - Publish it in a GitHub repository
  - Apply for a bot flag
  - Run the bot on a server



Wikidata:Requests for permissions/Bot/RefB (WikiCred)

< Wikidata:Requests for permissions | Bot

The following discussion is closed. ***Please do not modify it.*** *Subsequent comments should be made in a new section. A summary of the conclusions reached follows.*

✓ **Approved**--Ymblanter (talk) 10:10, 2 September 2020 (UTC)

RefB (WikiCred) [edit]

RefB (WikiCred) (talk • contribs • new items • SUL • Block log • User rights log • User rights • xtools)
Operator: Csisc (talk • contribs • logs)

Task/s: This bot will add reference support to biomedical statements in Wikidata.

Code: https://github.com/Data-Engineering-and-Semantics/refb/

Function details:

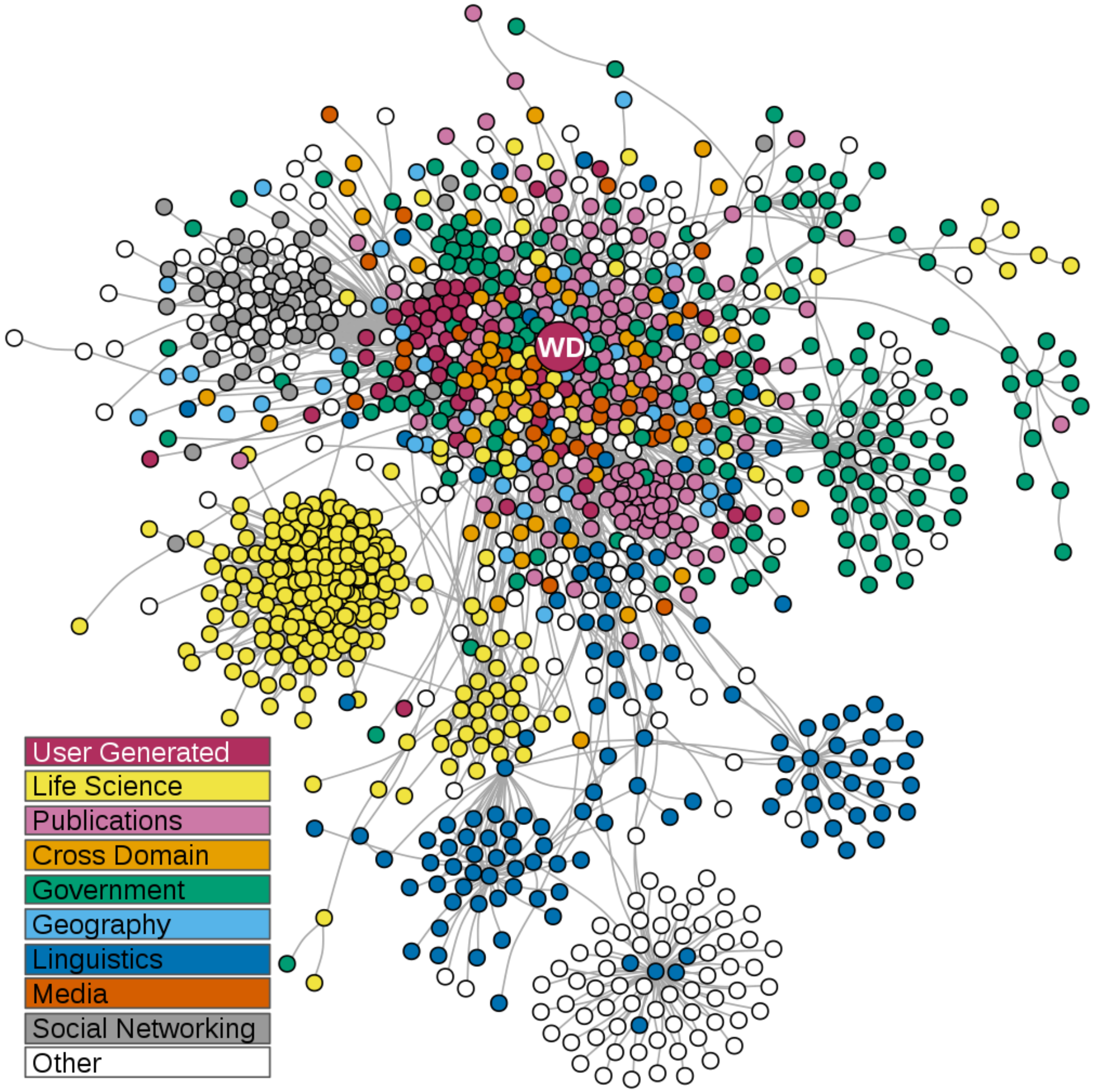
- This bot identifies unsupported biomedical relations on Wikidata using a SPARQL query.
- To find references supporting the extracted Wikidata statements, all that should be done is to use the PubMed Central search engine (based on Biopython and NCBI Entrez API) to find publications where the subject and the object of each statement co-occur. The algorithm will return the PMC ID of the reference for each assessed Wikidata statement and the sentence proving it within the full text of the reference.
- All we need to do is to convert PMC IDs into Wikidata IDs using Wikidata Hub, and then add the obtained references to Wikidata using the QuickStatements API.
- The source code of this bot is build using Python 3.5.
- Further details about the bot can be found here.

--Csisc (talk) 14:31, 29 July 2020 (UTC)

Adding COVID-19-related items and statements in KG

Task 3

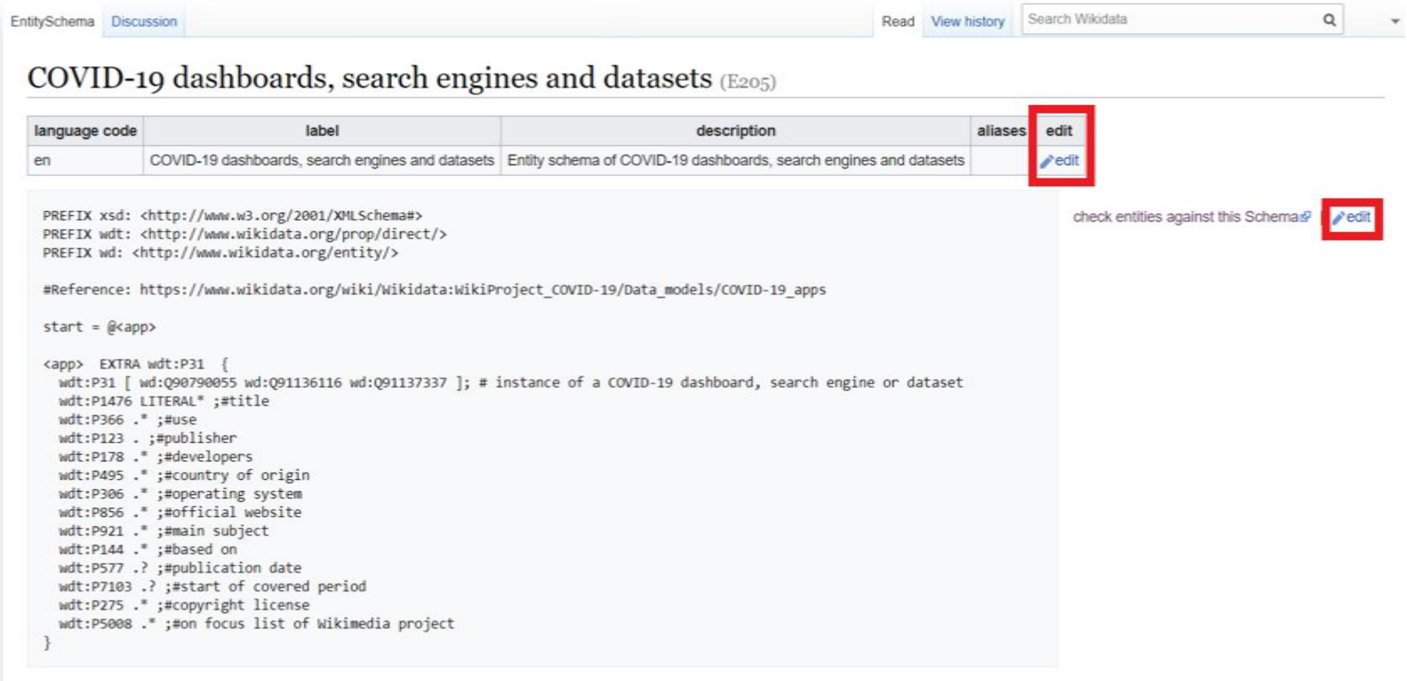
- Wikidata items are aligned to several open datasets and knowledge graphs particularly in the context of Linked Open Data Cloud
  - Other open knowledge graphs are automatically extracting COVID-19 information from CORD-19
  - Wikidata can integrate these information if the licenses of these open knowledge bases waive all the legal barriers (CC0 or Public Domain)



Updating the data model for COVID-19

Task 4

- Many methods:
  - Inferring new COVID-19 related classes from Topic Modelling of CORD-19 and mass import them to Wikidata using QuickStatements tool.
  - Deriving the data model of classes by analyzing the characteristics of COVID-19 knowledge in Wikidata using full screening and SPARQL queries.
  - Human updates of data models and ShEx validation schemas.



COVID-19 dashboards, search engines and datasets (E205)

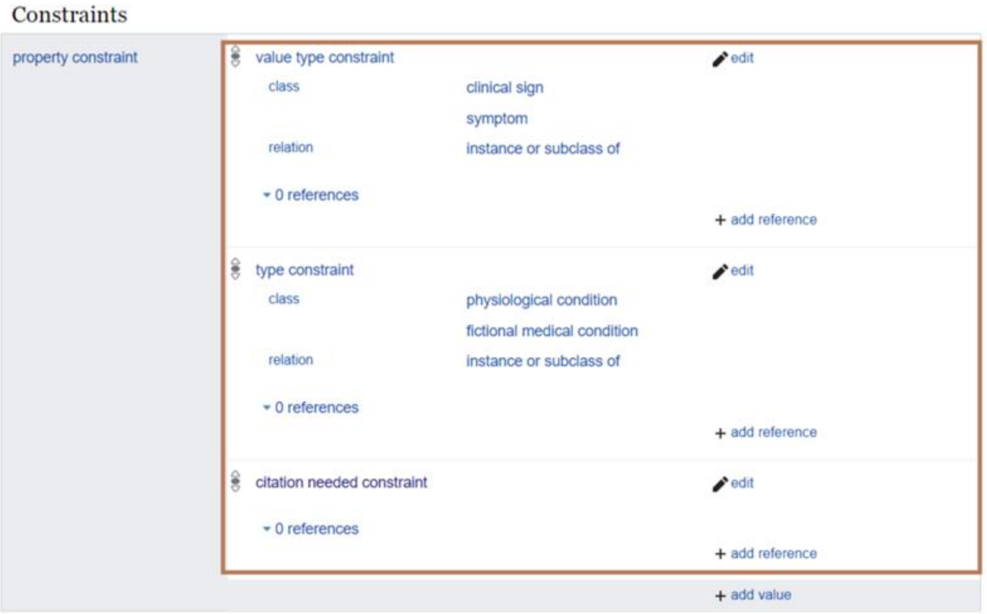
language code	label	description	aliases	edit
en	COVID-19 dashboards, search engines and datasets	Entity schema of COVID-19 dashboards, search engines and datasets		edit

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>

#Reference: https://www.wikidata.org/wiki/Wikidata:WikiProject_COVID-19/Data_models/COVID-19_apps

start = @<app>

<app>  EXTRA wdt:P31  {
  wdt:P31 [ wd:Q90790055 wd:Q91136116 wd:Q91137337 ]; # instance of a COVID-19 dashboard, search engine or dataset
  wdt:P1476 LITERAL* ;#title
  wdt:P366 .* ;#use
  wdt:P123 . ;#publisher
  wdt:P178 .* ;#developers
  wdt:P495 .* ;#country of origin
  wdt:P306 .* ;#operating system
  wdt:P856 .* ;#official website
  wdt:P921 .* ;#main subject
  wdt:P144 .* ;#based on
  wdt:P577 .? ;#publication date
  wdt:P7103 .? ;#start of covered period
  wdt:P275 .* ;#copyright license
  wdt:P5008 .* ;#on focus list of Wikimedia project
}
```

Validating existing COVID-19 informatio

Task 5

- Property constraints and statements
  - Define structural constraints for the definition of Wikidata statements
  - Identify the links between Wikidata properties

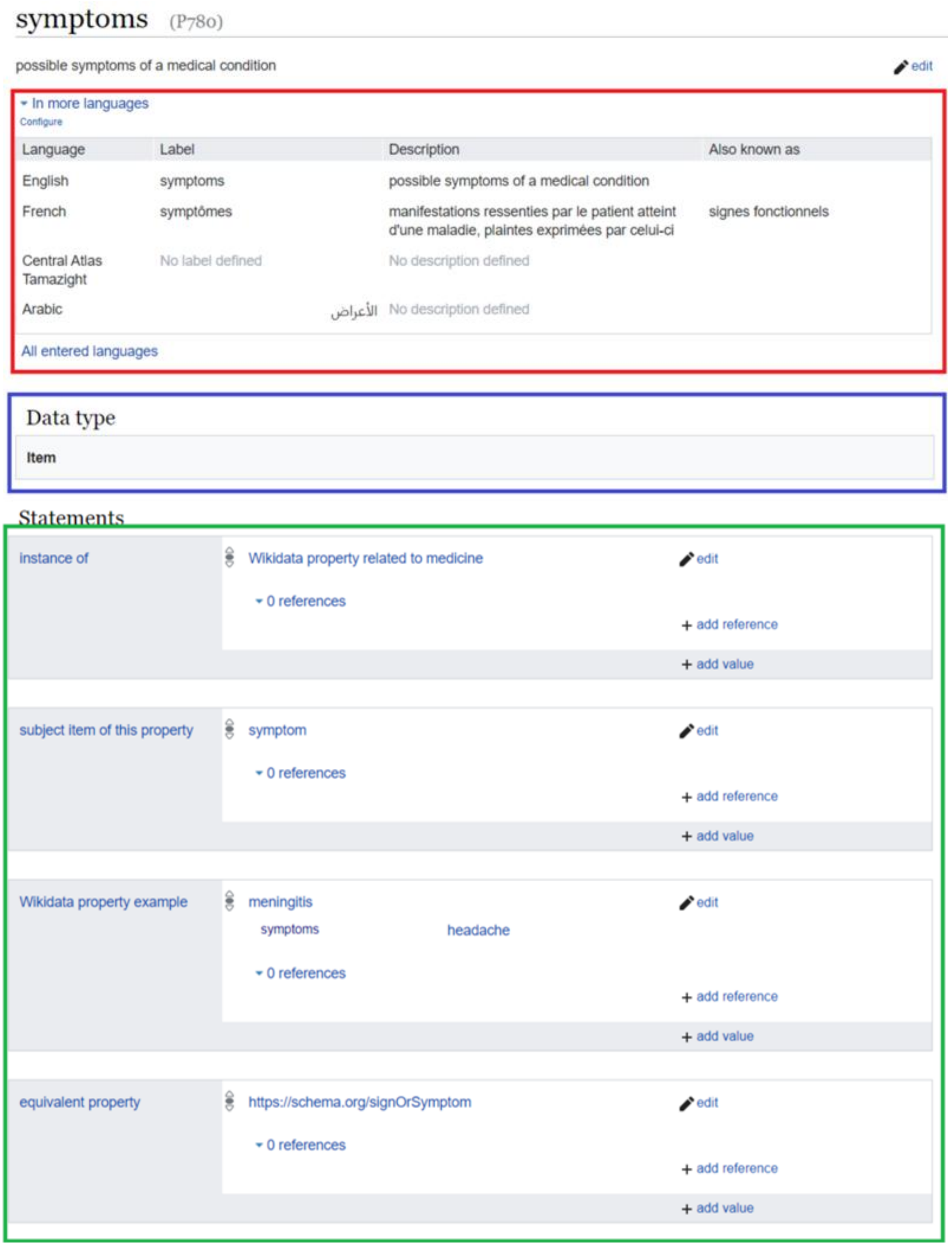


Constraints

property constraint

- value type constraint ✎ edit
  - class: clinical sign
symptom
  - relation: instance or subclass of
  - ▾ 0 references
  - + add reference
- type constraint ✎ edit
  - class: physiological condition
fictional medical condition
  - relation: instance or subclass of
  - ▾ 0 references
  - + add reference
- citation needed constraint ✎ edit
  - ▾ 0 references
  - + add reference

+ add value



symptoms (P780)

possible symptoms of a medical condition ✎ edit

▾ In more languages
Configure

Language	Label	Description	Also known as
English	symptoms	possible symptoms of a medical condition	
French	symptômes	manifestations ressenties par le patient atteint d'une maladie, plaintes exprimées par celui-ci	signes fonctionnels
Central Atlas Tamazight	No label defined	No description defined	
Arabic	الأعراض	No description defined	

All entered languages

Data type

Item

Statements

Property	Value	
instance of	Wikidata property related to medicine ▾ 0 references + add reference	✎ edit + add value
subject item of this property	symptom ▾ 0 references + add reference	✎ edit + add value
Wikidata property example	meningitis symptoms: headache ▾ 0 references + add reference	✎ edit + add value
equivalent property	https://schema.org/signOrSymptom ▾ 0 references + add reference	✎ edit + add value

Validating existing COVID-19 information

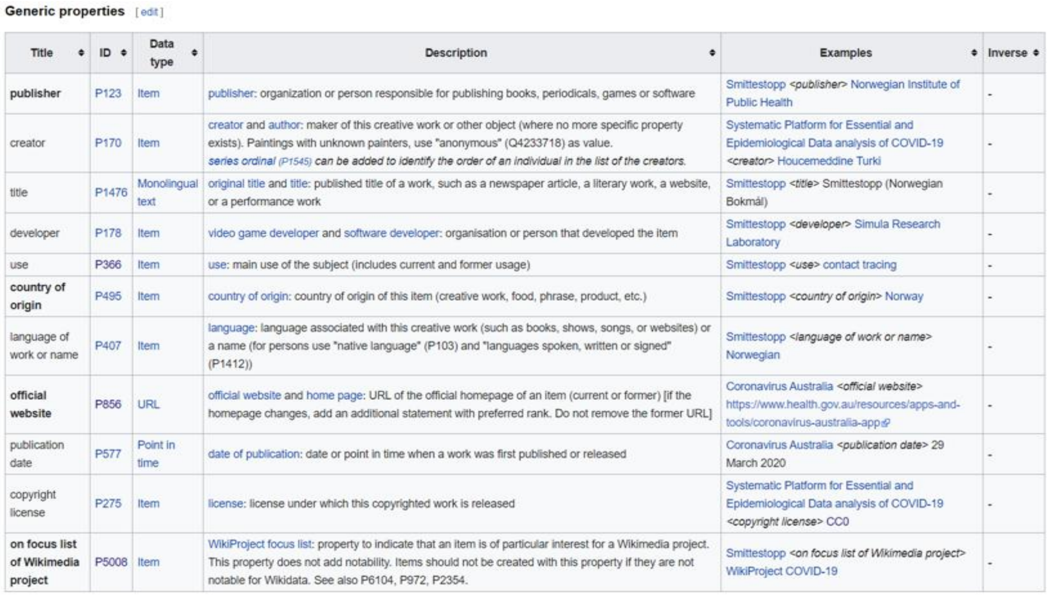
Task 5

Constraint type	Description
single value constraint	Constraint used to specify that this property generally contains a single value per item
format constraint	Constraint used to specify that the value for this property has to correspond to a given pattern
mandatory constraint	status of a Wikidata property constraint: indicates that the specified constraint applies to the subject property without exception and must not be violated
distinct values constraint	Constraint used to specify that the value for this property is likely to be different from all other items
Commons link constraint	Constraint used to specify that the value must link to an existing Wikimedia Commons page
difference within range constraint	Constraint used to specify that the value of a given statement should only differ in the given way. Use with qualifiers minimum quantity/maximum quantity
mandatory qualifier constraint	Constraint used to specify that the listed qualifier has to be used
symmetric constraint	Constraint used to specify that the referenced entity should also link back to this entity
used as qualifier constraint	Constraint used to specify that a property must only be used as a qualifier
value requires statement constraint	Constraint used to specify that the referenced item should have a statement with a given property
relation of type constraint	relation establishing dependency between types/metalevels of its members
allowed qualifiers constraint	Constraint used to specify that only the listed qualifiers should be used. Novalue disallows any qualifier
value type constraint	Constraint used to specify that the referenced item should be a subclass or instance of a given type
allowed units constraint	Constraint used to specify that only listed units may be used
multi-value constraint	Constraint used to specify that a property generally contains more than one value per item
one-of constraint	Constraint used to specify that the value for this property has to be one of a given set of items
range constraint	Constraint used to specify that the value must be between two given values

Validating existing COVID-19 information

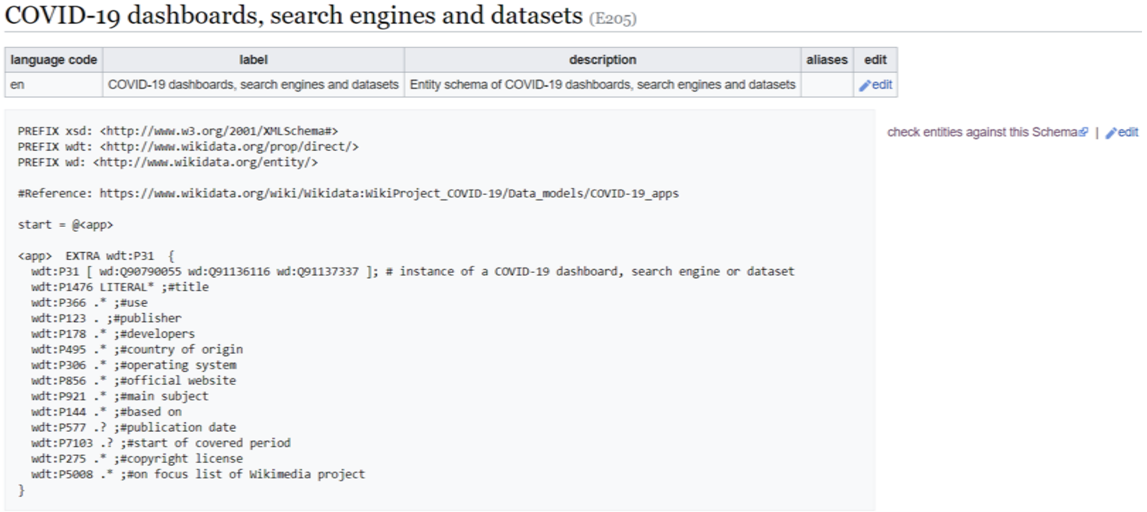
Task 5

- Data Models and ShEx
  - Specify the required statements for the definition of a Wikidata item
  - Available at https://www.wikidata.org/wiki/Wikidata:WikiProject_COVID-19/Data_models



Generic properties [edit]

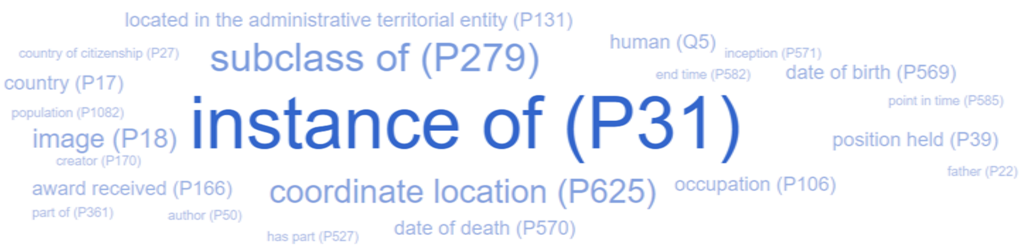
Title	ID	Data type	Description	Examples	Inverse
publisher	P123	Item	publisher: organization or person responsible for publishing books, periodicals, games or software	Smittestopp *<publisher>* Norwegian Institute of Public Health	-
creator	P170	Item	creator and author: maker of this creative work or other object (where no more specific property exists). Paintings with unknown painters, use "anonymous" (Q4233718) as value. *series ordinal (P1545) can be added to identify the order of an individual in the list of the creators.*	Systematic Platform for Essential and Epidemiological Data analysis of COVID-19 *<creator>* Houcemeddine Turki	-
title	P1476	Monolingual text	original title and title: published title of a work, such as a newspaper article, a literary work, a website, or a performance work	Smittestopp *<title>* Smittestopp (Norwegian Bokmål)	-
developer	P178	Item	video game developer and software developer: organisation or person that developed the item	Smittestopp *<developer>* Simula Research Laboratory	-
use	P366	Item	use: main use of the subject (includes current and former usage)	Smittestopp *<use>* contact tracing	-
country of origin	P495	Item	country of origin: country of origin of this item (creative work, food, phrase, product, etc.)	Smittestopp *<country of origin>* Norway	-
language of work or name	P407	Item	language: language associated with this creative work (such as books, shows, songs, or websites) or a name (for persons use "native language" (P103) and "languages spoken, written or signed" (P1412))	Smittestopp *<language of work or name>* Norwegian	-
official website	P856	URL	official website and home page: URL of the official homepage of an item (current or former) [if the homepage changes, add an additional statement with preferred rank. Do not remove the former URL]	Coronavirus Australia *<official website>* https://www.health.gov.au/resources/apps-and-tools/coronavirus-australia-app	-
publication date	P577	Point in time	date of publication: date or point in time when a work was first published or released	Coronavirus Australia *<publication date>* 29 March 2020	-
copyright license	P275	Item	license: license under which this copyrighted work is released	Systematic Platform for Essential and Epidemiological Data analysis of COVID-19 *<copyright license>* CC0	-
on focus list of Wikimedia project	P5008	Item	WikiProject focus list: property to indicate that an item is of particular interest for a Wikimedia project. This property does not add notability. Items should not be created with this property if they are not notable for Wikidata. See also P6104, P972, P2354.	Smittestopp *<on focus list of Wikimedia project>* WikiProject COVID-19	-

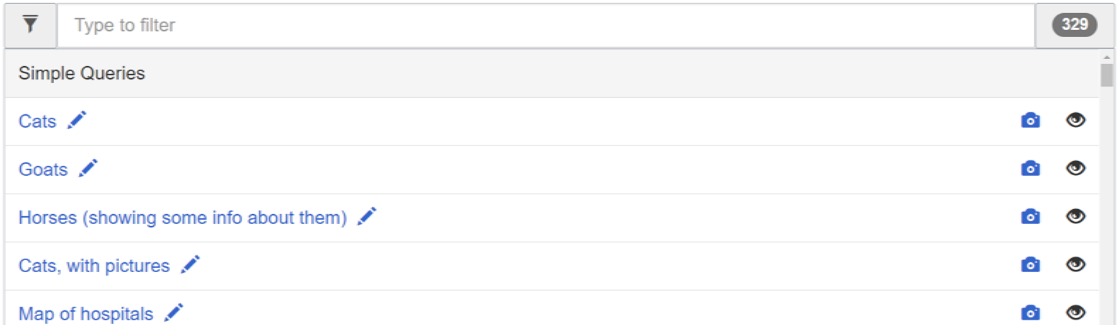


COVID-19 dashboards, search engines and datasets (E205)

language code	label	description	aliases	edit
en	COVID-19 dashboards, search engines and datasets	Entity schema of COVID-19 dashboards, search engines and datasets		edit

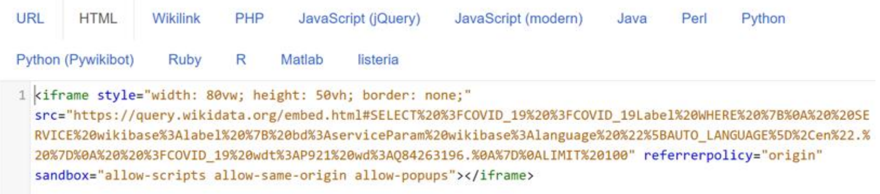
```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>

#Reference: https://www.wikidata.org/wiki/Wikidata:WikiProject_COVID-19/Data_models/COVID-19_apps

start = @<app>

<app>  EXTRA wdt:P31  {
  wdt:P31 [ wd:Q90790055 wd:Q91136116 wd:Q91137337 ]; # instance of a COVID-19 dashboard, search engine or dataset
  wdt:P1476 LITERAL* ;#title
  wdt:P366 .* ;#use
  wdt:P123 . ;#publisher
  wdt:P178 .* ;#developers
  wdt:P495 .* ;#country of origin
  wdt:P306 .* ;#operating system
  wdt:P856 .* ;#official website
  wdt:P921 .* ;#main subject
  wdt:P144 .* ;#based on
  wdt:P577 .? ;#publication date
  wdt:P7103 .? ;#start of covered period
  wdt:P275 .* ;#copyright license
  wdt:P5008 .* ;#on focus list of Wikimedia project
}
```



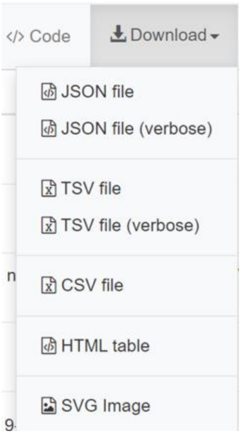


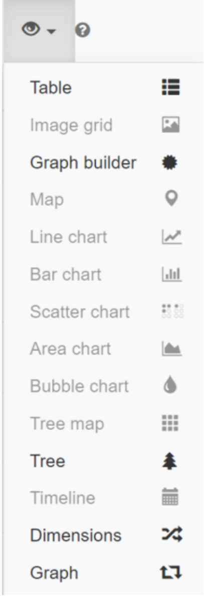
Simple Queries

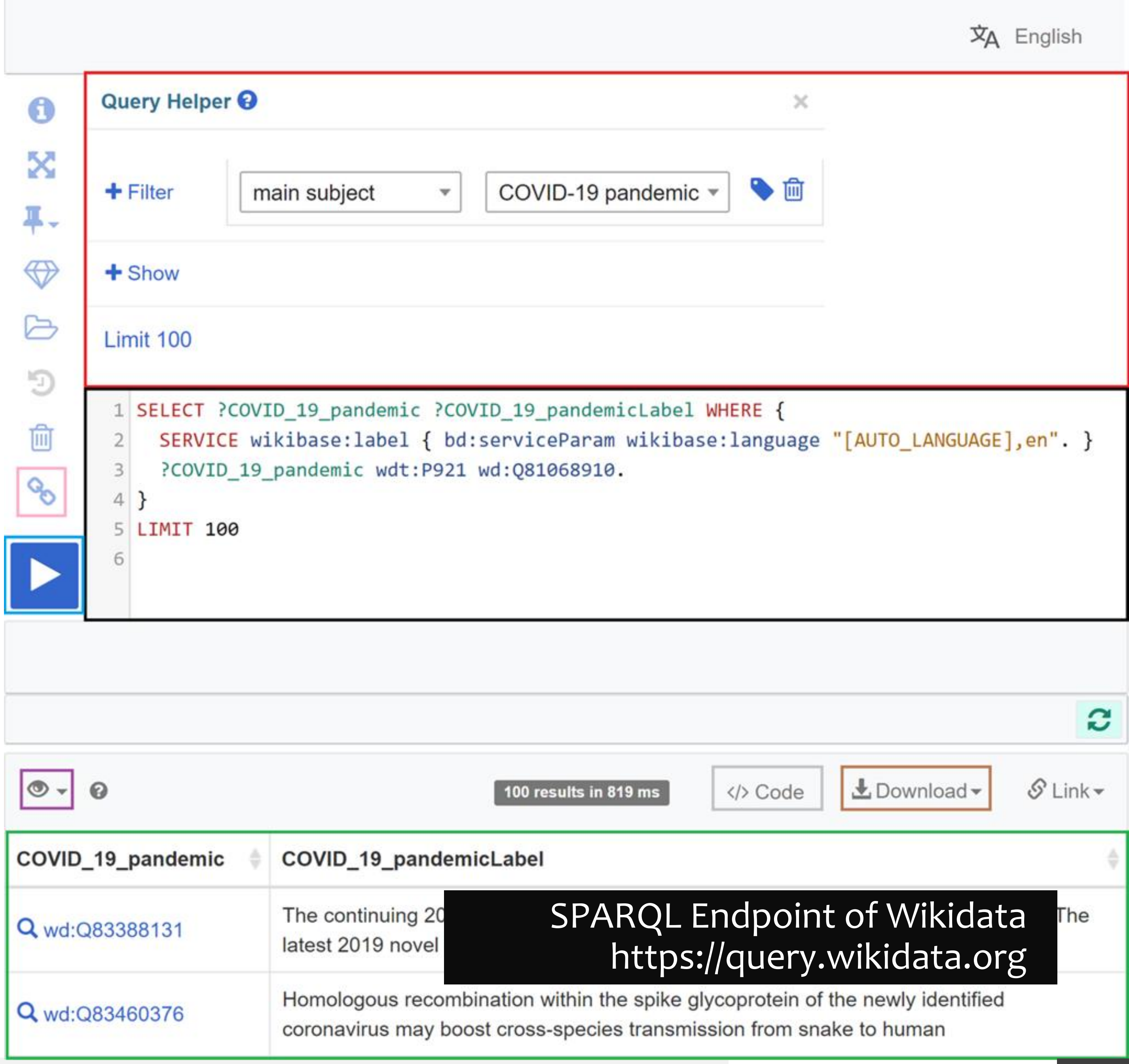
- Cats
- Goats
- Horses (showing some info about them)
- Cats, with pictures
- Map of hospitals



```
<iframe style="width: 80vw; height: 50vh; border: none;" src="https://query.wikidata.org/embed.html#SELECT%20%3FCOVID_19%20%3FCOVID_19Label%20WHERE%20%7B%0A%20%20SERVICE%20wikibase%3Alabel%20%7B%20bd%3AserviceParam%20wikibase%3Alanguage%20%22%5BAUTO_LANGUAGE%5D%2Cen%22.%20%7D%0A%20%20%3FCOVID_19%20wdt%3AP921%20wd%3AQ84263196.%0A%7D%0ALIMIT%20100" referrerpolicy="origin" sandbox="allow-scripts allow-same-origin allow-popups"></iframe>
```







```
SELECT ?COVID_19_pandemic ?COVID_19_pandemicLabel WHERE {
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
  ?COVID_19_pandemic wdt:P921 wd:Q81068910.
}
LIMIT 100
```

COVID_19_pandemic	COVID_19_pandemicLabel
wd:Q83388131	The continuing 20 … The latest 2019 novel
wd:Q83460376	Homologous recombination within the spike glycoprotein of the newly identified coronavirus may boost cross-species transmission from snake to human

SPARQL Endpoint of Wikidata
https://query.wikidata.org

Validating existing COVID-19 information

Task 5

- Logical constraints implemented in SPARQL to validate relational statements

Constraint	Description
Defining the scheme of a Wikidata property	
T1	Identify common use cases of *P*: (C_S,C_O) pairs
T2	Identify inverse properties of *P* corresponding to each common use case: (C_S, R^{-1},C_O) statements
Identifying the deficiencies of the scheme	
T3	For each returned P^{-1}, identify *P*(*S*,*O*) relations supported by references and corresponding to the most common (C_S, P^{-1}, C_O) statement but not available in Wikidata
T4	Identify *P*(*S*,*O*) relations not corresponding to the most common scheme of *P*

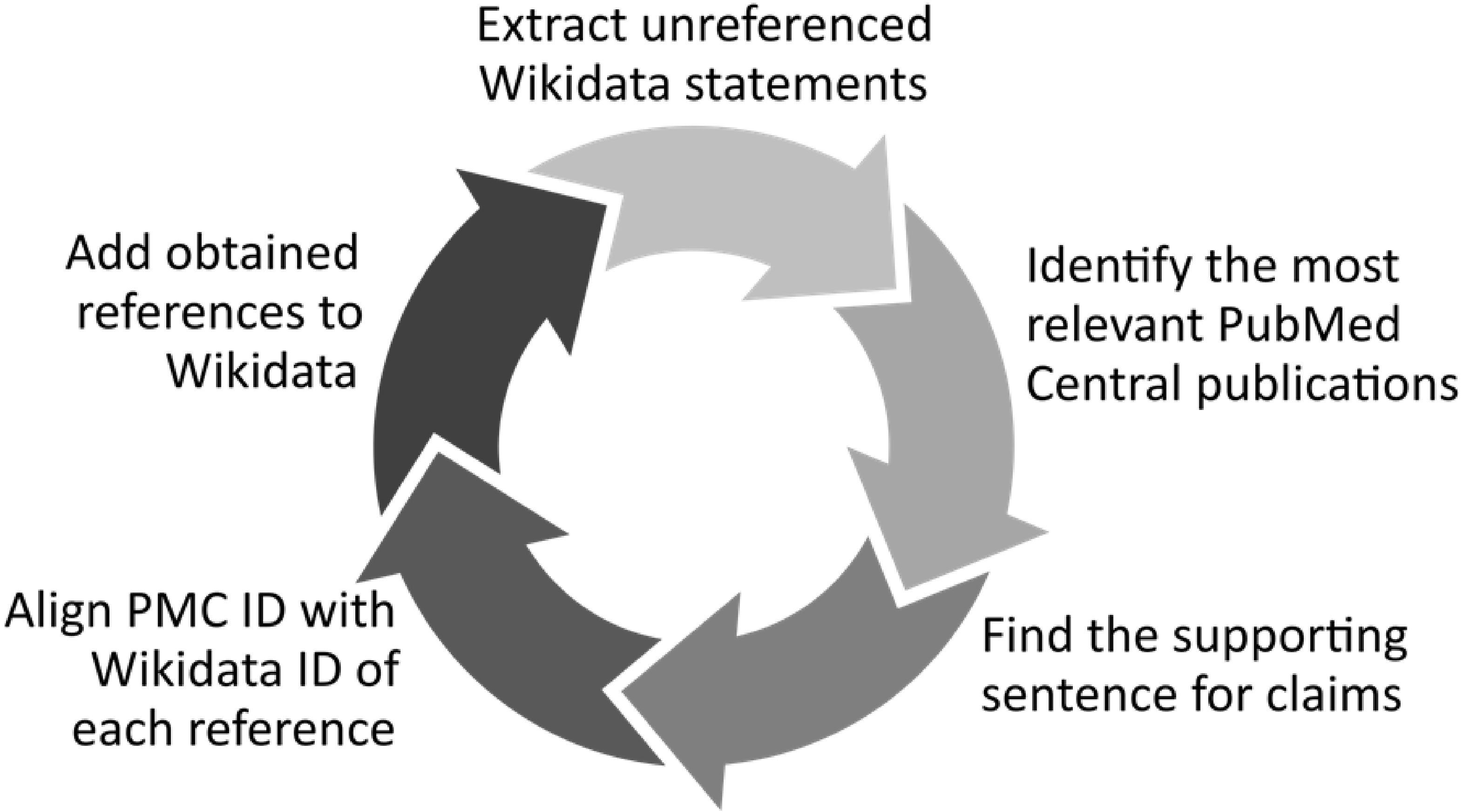
- Logical constraints implemented in SPARQL to validate statistical statements

Constraint	Description
Validating qualifiers of COVID-19 epidemiological statements	
V1	Verify *Z* as a date > November 01, 2019
V2	Verify *Q* as any subclass of (P279*) of medical diagnosis (Q177719)
Ensuring the cumulative pattern of *c*, *d*, *r*, and *t*	
V3	Identify *c*, *d*, *r* and *t* statements having a value in date *Z*+1 not superior or equal to the one in date *Z* (Verify if $d_Z \leq d_{Z+1}$, $r_Z \leq r_{Z+1}$, $t_Z \leq t_{Z+1}$, and $c_Z \leq c_{Z+1}$)
V4	Find missing values of *c*, *d*, *r* and *t* in date *Z*+1 where corresponding values in dates *Z* and *Z*+2 are equal

Validating existing COVID-19 information

Task 5

- Scholarly databases like CORD-19 should not only be used to extract COVID-19 information.
- Scholarly databases can be searched to find interesting references for unsupported statements in Wikidata



Extract unreferenced Wikidata statements

Identify the most relevant PubMed Central publications

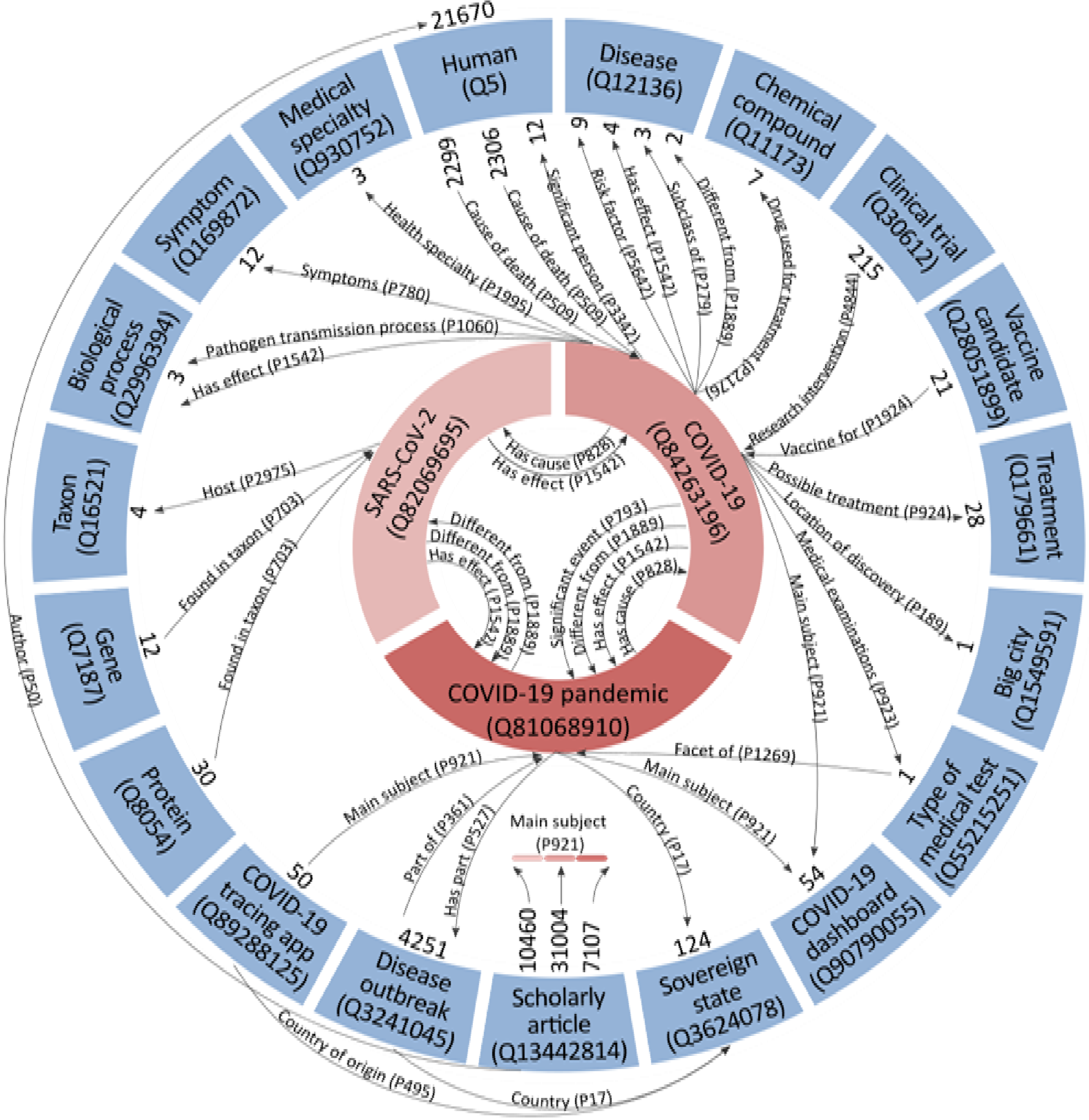
Find the supporting sentence for claims

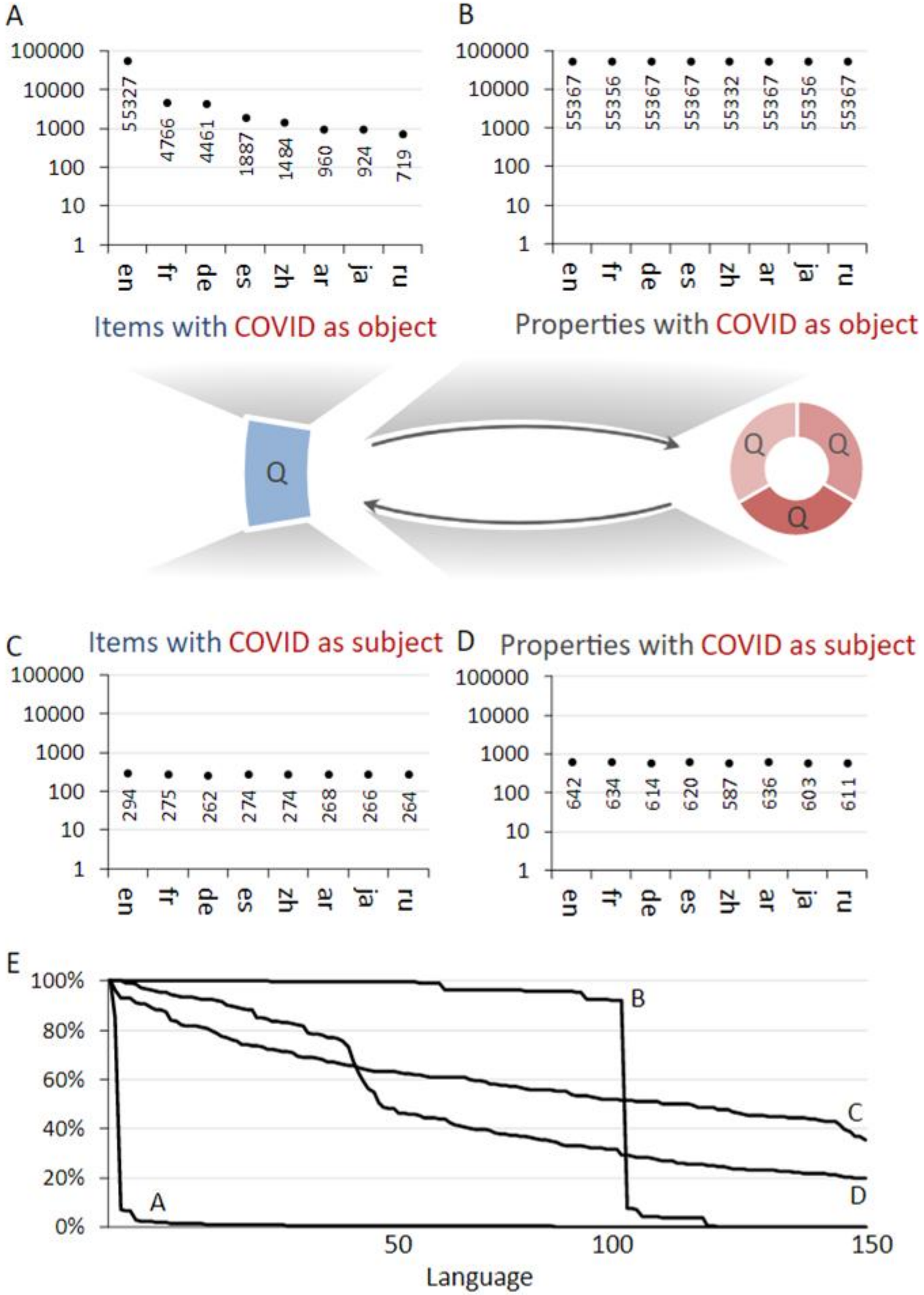
Align PMC ID with Wikidata ID of each reference

Add obtained references to Wikidata

Results and Discussion

Current situation of COVID-19 semantic information in Wikidata





Interesting coverage of multiple facets of COVID-19 information

Add a footer

Factors for the growth of COVID-19 multilingual coverage in Wikidata

Positive correlations between the language support for COVID-19 and a significant number of factors

	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

	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

External databases aligned to Wikidata items

Scholarly research publications and clinical trials

Wikidata ID	Wikidata Property	Count
P356	DOI	45101
P698	PubMed ID	42294
P6179	Dimensions Publication ID	16944
P932	PMCID	12590

Diseases and clinical signs

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	32	1
P3827	JSTOR topic ID	32	10

External databases aligned to Wikidata items

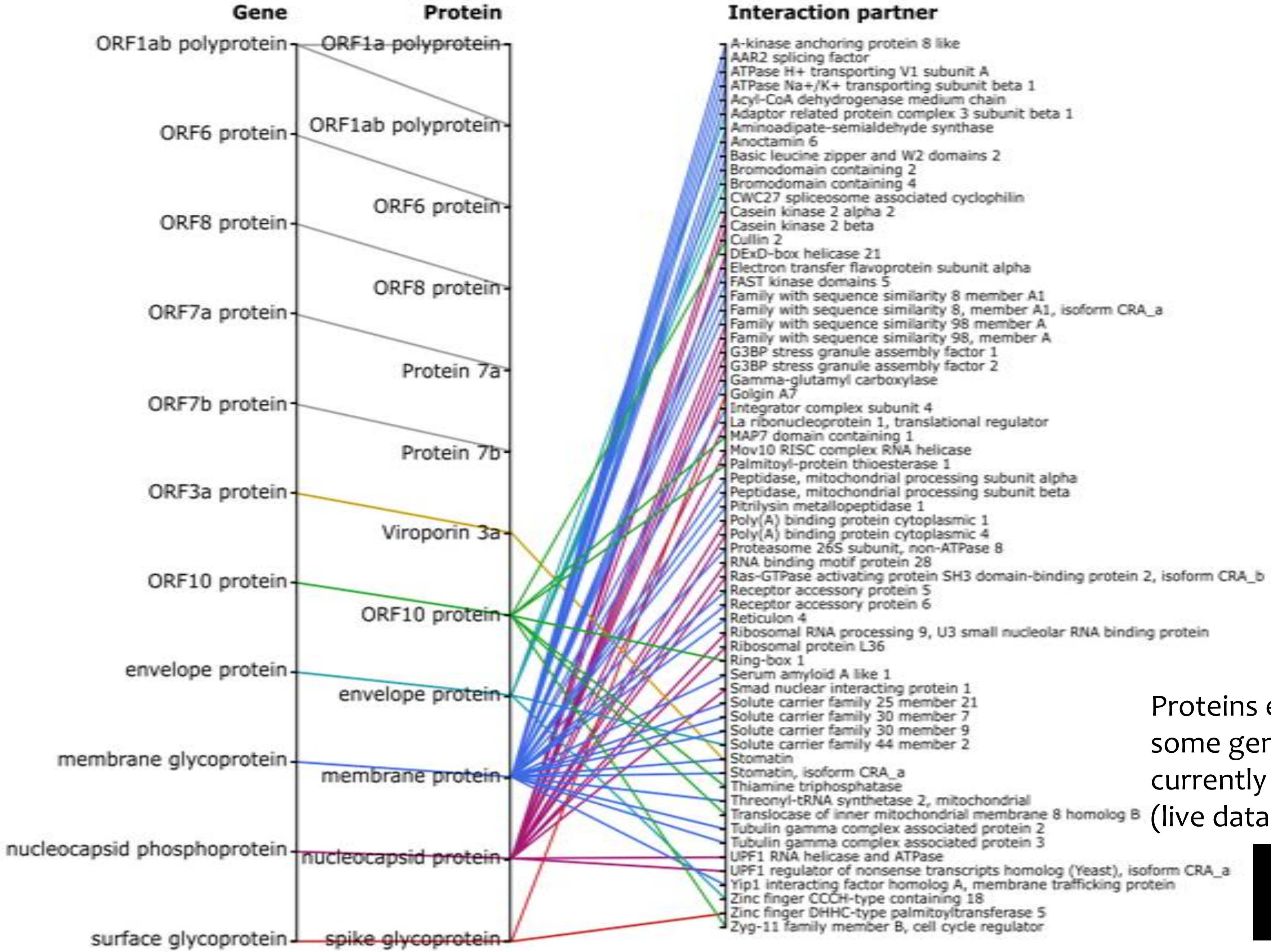
Humans and sovereign states

Wikidata ID	Wikidata Property	Sovereign states	Humans
P214	VIAF ID	159	654
P7859	WorldCat Identities ID	146	548
P244	Library of Congress authority ID	125	458
P213	ISNI	100	443
P646	Freebase ID	124	379
P2002	Twitter username	16	353

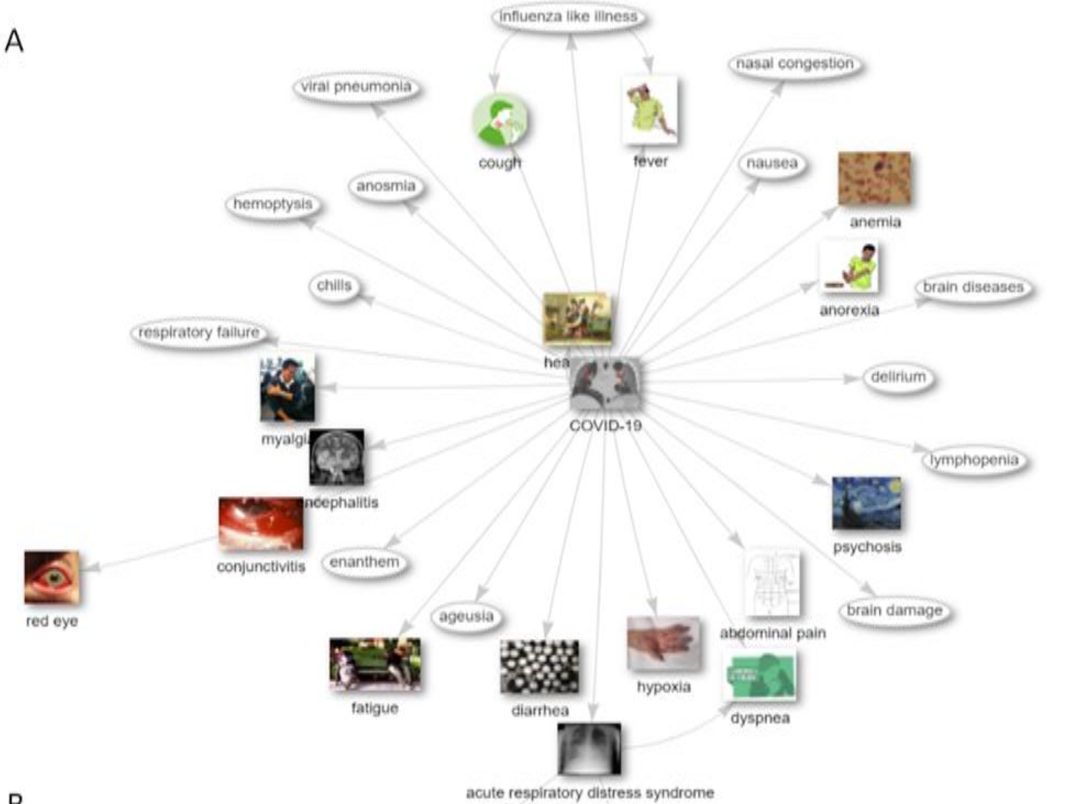
Other items

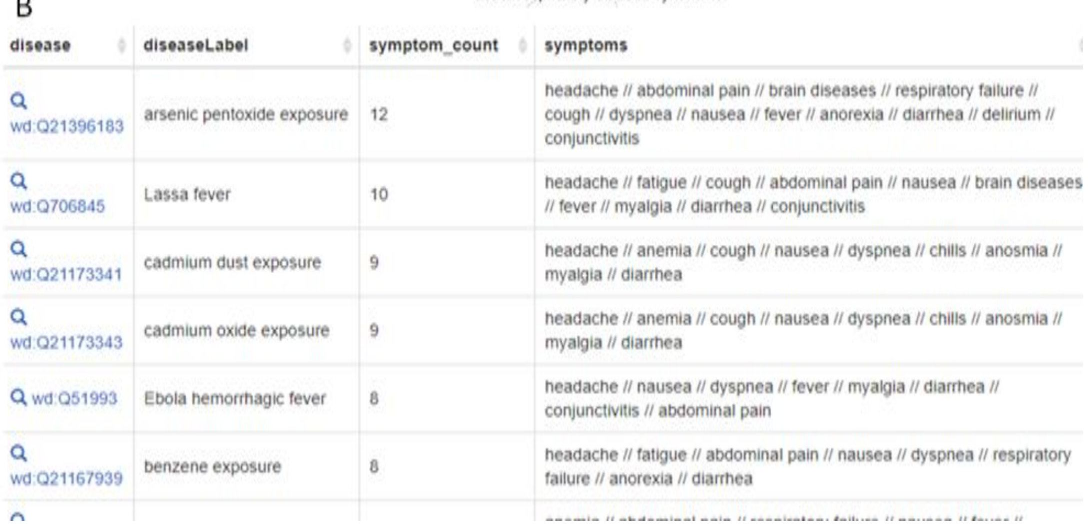
Wikidata Class	Wikidata ID	Wikidata Property	Count
drug [Q11173]	P6689	MassBank accession ID	44
drug [Q11173]	P4964	SPLASH	31
protein [Q8054]	P638	PDB structure ID	31
film [Q11424]	P345	IMDb ID	25

SARS-CoV-2 interactions with the human proteome


Gene
Protein
Interaction partner
ORF1ab polyprotein
ORF6 protein
ORF8 protein
ORF7a protein
ORF7b protein
ORF3a protein
ORF10 protein
envelope protein
membrane glycoprotein
nucleocapsid phosphoprotein
surface glycoprotein
ORF1a polyprotein
ORF1ab polyprotein
ORF6 protein
ORF8 protein
Protein 7a
Protein 7b
Viroporin 3a
ORF10 protein
envelope protein
membrane protein
nucleocapsid protein
spike glycoprotein
A-kinase anchoring protein 8 like
AAR2 splicing factor
ATPase H+ transporting V1 subunit A
ATPase Na+/K+ transporting subunit beta 1
Acyl-CoA dehydrogenase medium chain
Adaptor related protein complex 3 subunit beta 1
Aminoadipate-semialdehyde synthase
Anoctamin 6
Basic leucine zipper and W2 domains 2
Bromodomain containing 2
Bromodomain containing 4
CWC27 spliceosome associated cyclophilin
Casein kinase 2 alpha 2
Casein kinase 2 beta
Cullin 2
DExD-box helicase 21
Electron transfer flavoprotein subunit alpha
FAST kinase domains 5
Family with sequence similarity 8 member A1
Family with sequence similarity 8, member A1, isoform CRA_a
Family with sequence similarity 98 member A
Family with sequence similarity 98, member A
G3BP stress granule assembly factor 1
G3BP stress granule assembly factor 2
Gamma-glutamyl carboxylase
Golgin A7
Integrator complex subunit 4
La ribonucleoprotein 1, translational regulator
MAP7 domain containing 1
Mov10 RISC complex RNA helicase
Palmitoyl-protein thioesterase 1
Peptidase, mitochondrial processing subunit alpha
Peptidase, mitochondrial processing subunit beta
Pitrilysin metallopeptidase 1
Poly(A) binding protein cytoplasmic 1
Poly(A) binding protein cytoplasmic 4
Proteasome 26S subunit, non-ATPase 8
RNA binding motif protein 28
Ras-GTPase activating protein SH3 domain-binding protein 2, isoform CRA_b
Receptor accessory protein 5
Receptor accessory protein 6
Reticulon 4
Ribosomal RNA processing 9, U3 small nucleolar RNA binding protein
Ribosomal protein L36
Ring-box 1
Serum amyloid A like 1
Smad nuclear interacting protein 1
Solute carrier family 25 member 21
Solute carrier family 30 member 7
Solute carrier family 30 member 9
Solute carrier family 44 member 2
Stomatin
Stomatin, isoform CRA_a
Thiamine triphosphatase
Threonyl-tRNA synthetase 2, mitochondrial
Translocase of inner mitochondrial membrane 8 homolog B
Tubulin gamma complex associated protein 2
Tubulin gamma complex associated protein 3
UPF1 RNA helicase and ATPase
UPF1 regulator of nonsense transcripts homolog (Yeast), isoform CRA_a
Yip1 interacting factor homolog A, membrane trafficking protein
Zinc finger CCCH-type containing 18
Zinc finger DHHC-type palmitoyltransferase 5
Zyg-11 family member B, cell cycle regulator

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).





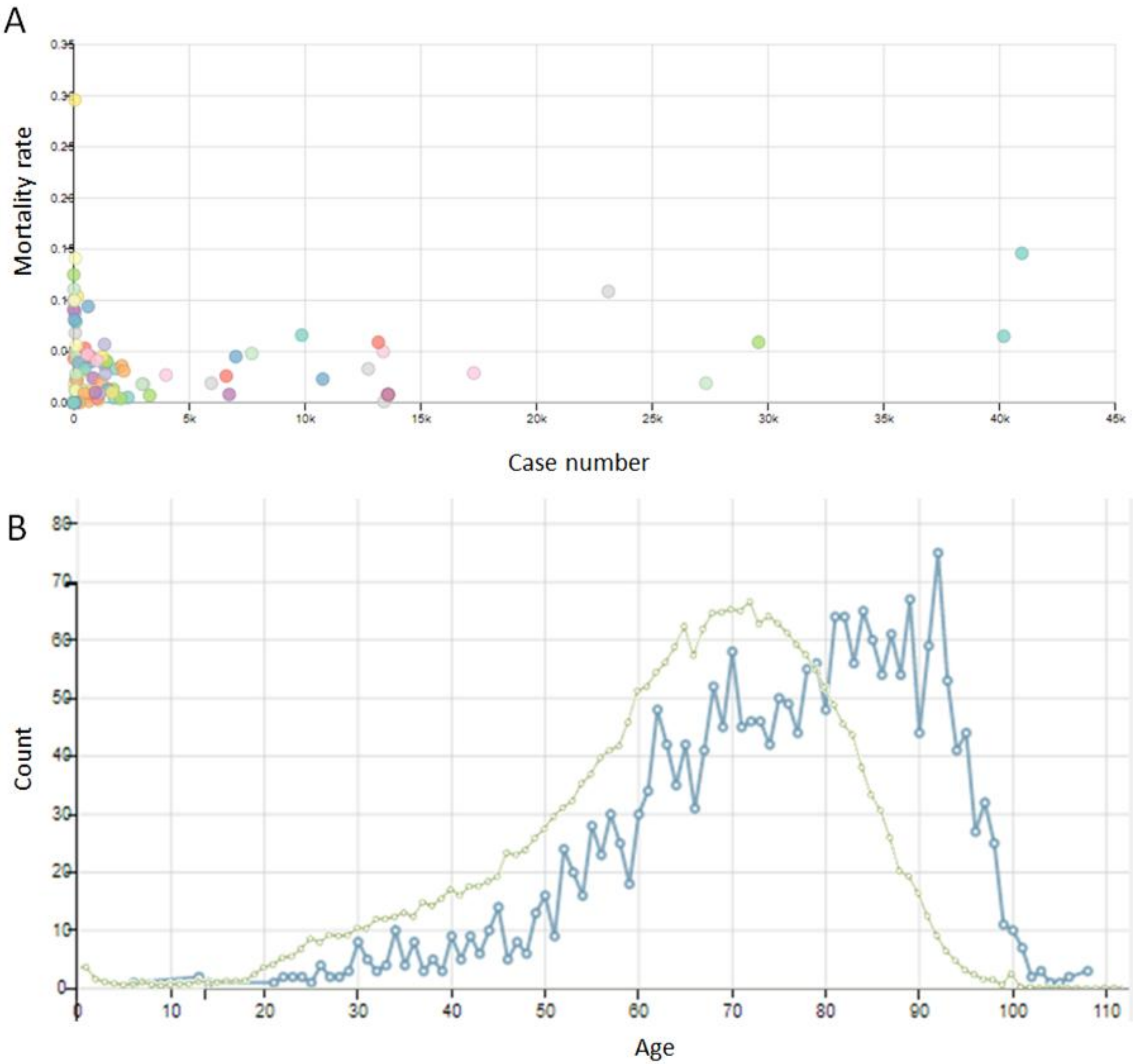
disease	diseaseLabel	symptom_count	symptoms
wd:Q21396183	arsenic pentoxide exposure	12	headache // abdominal pain // brain diseases // respiratory failure // cough // dyspnea // nausea // fever // anorexia // diarrhea // delirium // conjunctivitis
wd:Q706845	Lassa fever	10	headache // fatigue // cough // 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
			anemia // abdominal pain // respiratory failure // nausea // fever //

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)

Symptoms of COVID-19 and similar conditions

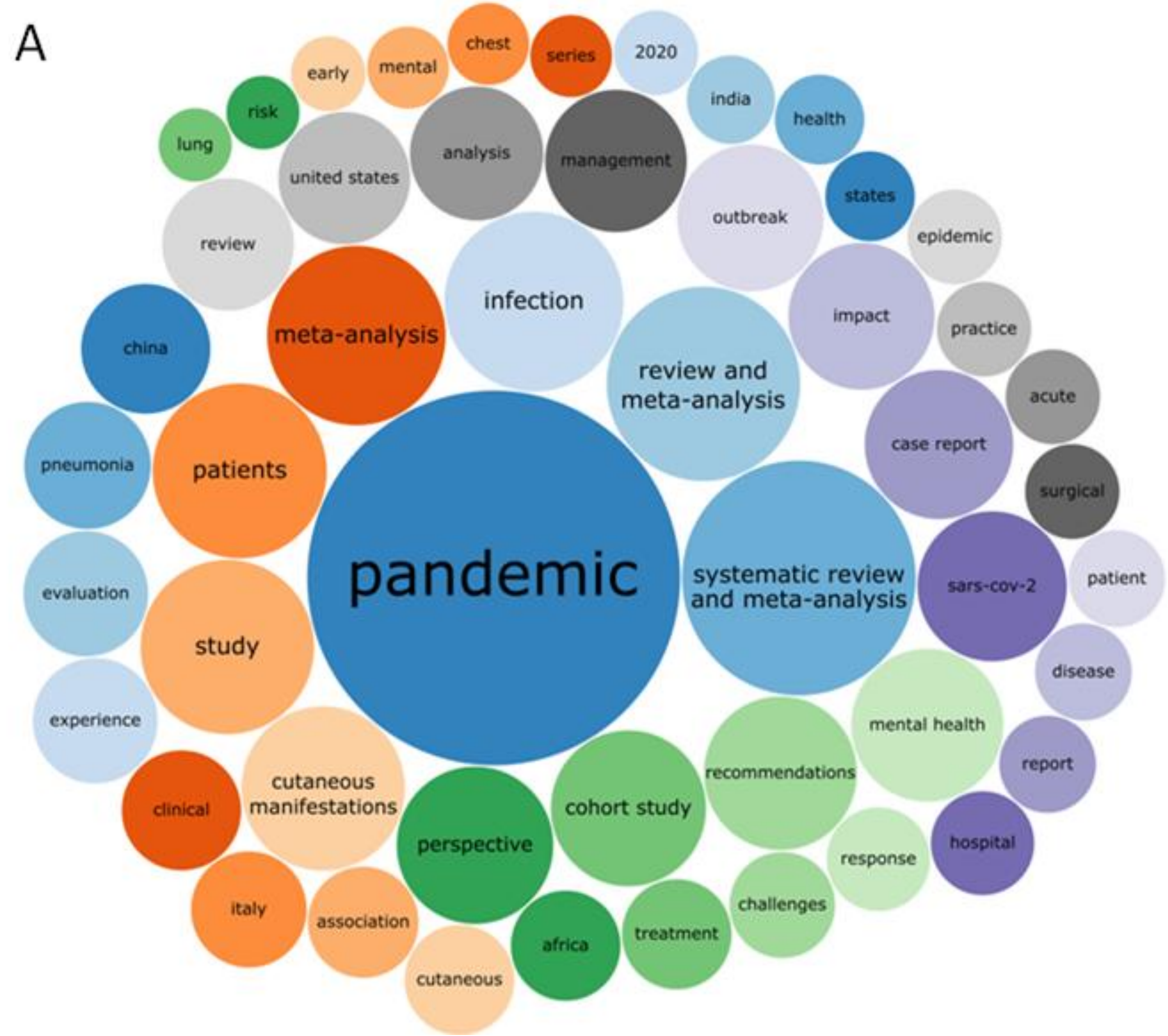
Summary epidemiological data on the COVID-19 pandemic

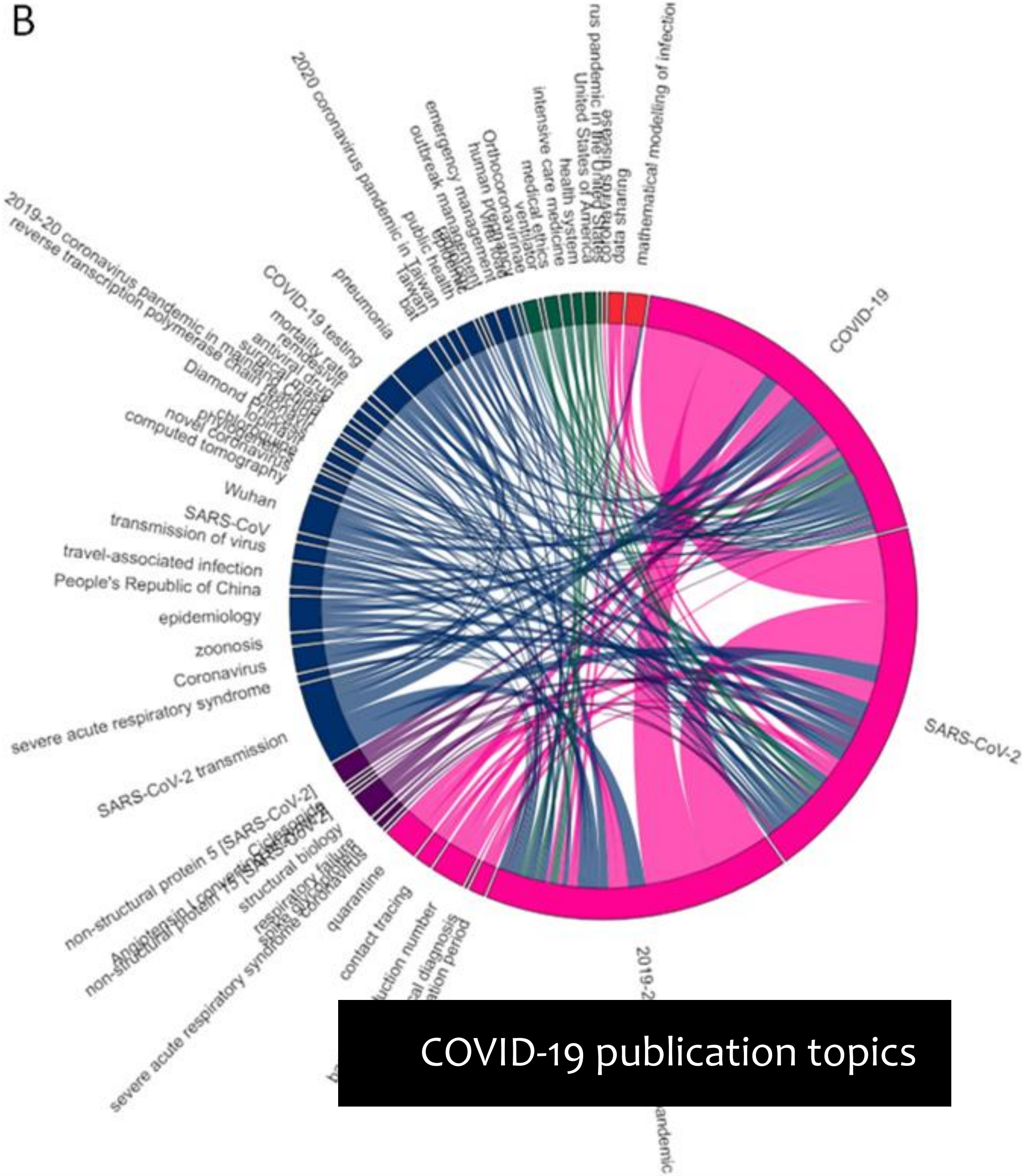


A) Correlation between the current number of cases and mortality rates in every country, calculated from numeric summary data for each region. Countries coloured randomly (live data: https://w.wiki/bf$).

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).

COVID-19 publication topics

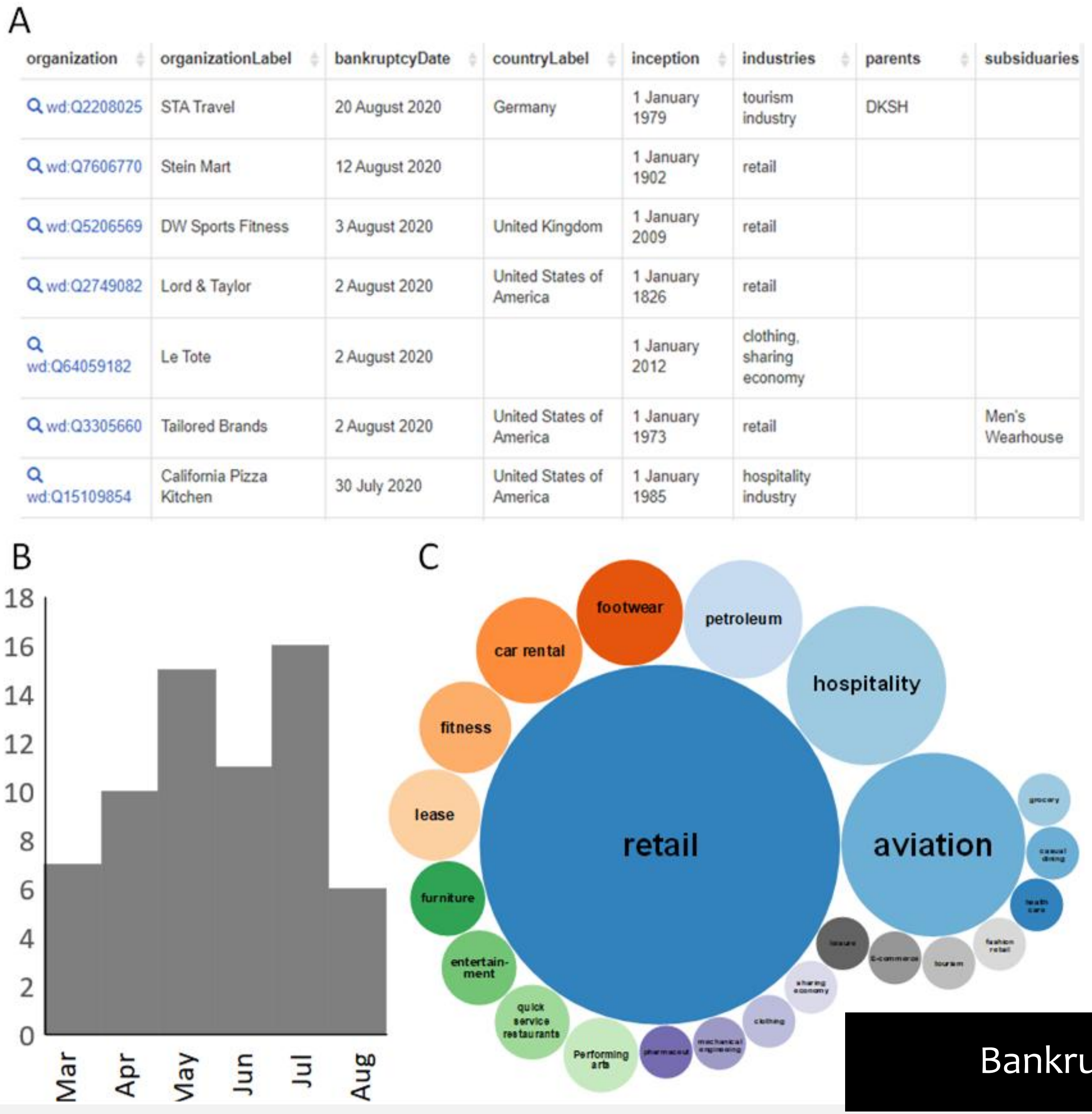




COVID-19 publication topics

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).



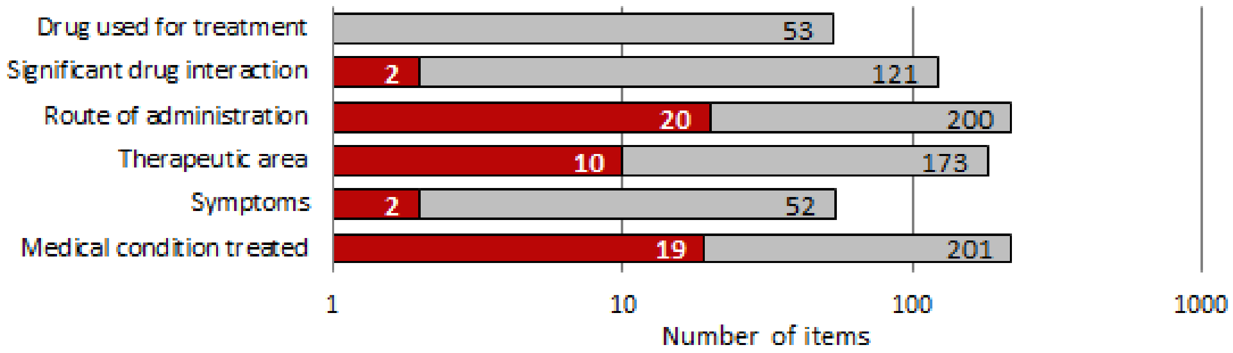
organization	organizationLabel	bankruptcyDate	countryLabel	inception	industries	parents	subsiduaries
wd:Q2208025	STA Travel	20 August 2020	Germany	1 January 1979	tourism industry	DKSH	
wd:Q7606770	Stein Mart	12 August 2020		1 January 1902	retail		
wd:Q5206569	DW Sports Fitness	3 August 2020	United Kingdom	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		

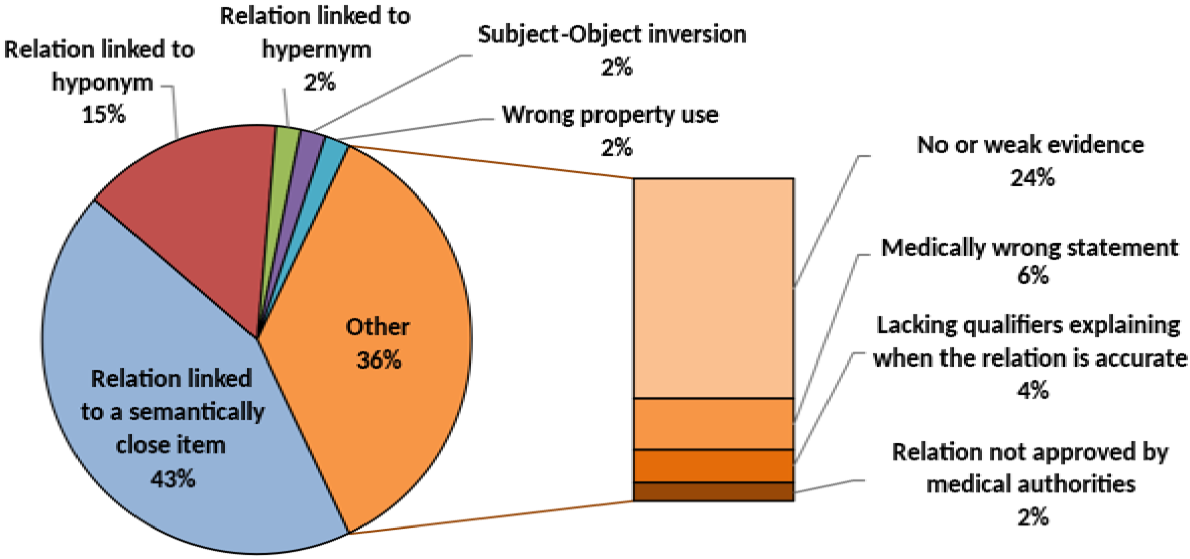
A) Tabular output of SPARQL query
B) Bankruptcies per month
C) ratios of different industries associated with bankrupt companies. (live data: https://w.wiki/cG6).

Bankrupt publicly listed businesses due to the COVID-19 pandemic

Validation using logical constraints

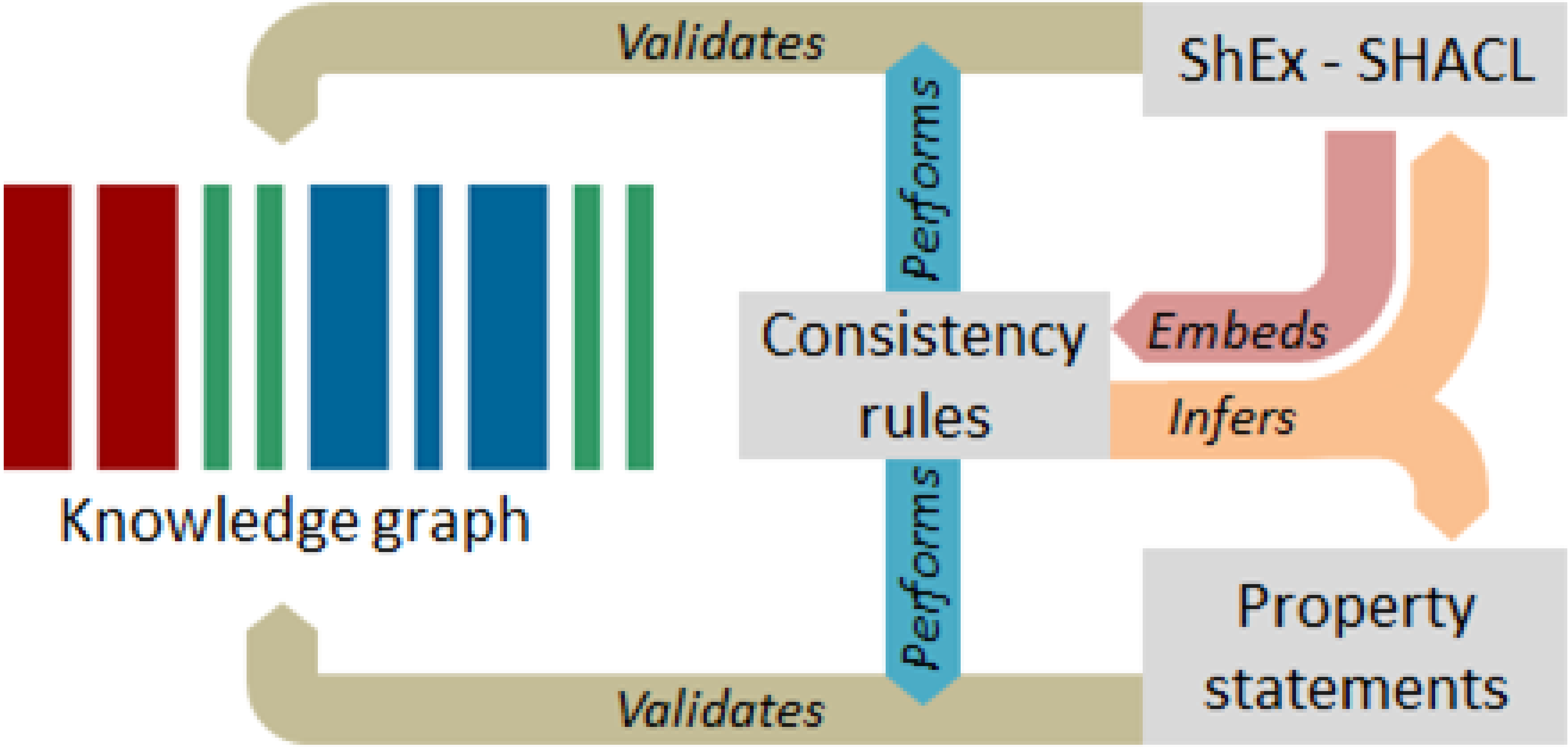
Relational statements





Statistical statements

cases	*deaths*	*recoveries*	*tests*	*hospitalizations*	**Overall**
2856	2467	189	9	10	5496



An infrastructure for knowledge graph validation based on interactions between consistency rules, property statements and RDF validation languages

Limitations

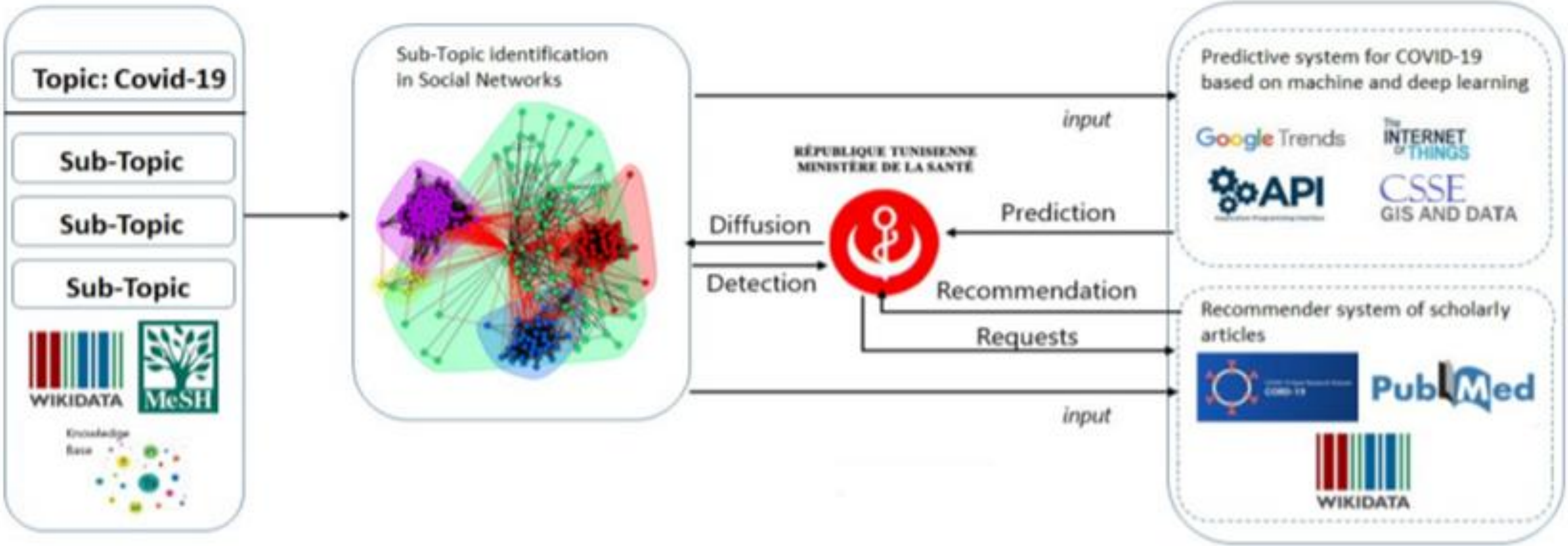
- Several aspects of COVID-19 information can be more represented in other knowledge graphs. The inclusion of these COVID-19 information is blocked by the CC0 License of Wikidata (e.g. ORKG)
- Several types of information are still not supported by Wikidata (e.g. Structured outcomes of COVID-19 scholarly publications)
- Several aspects of COVID-19 information are more considered in textual resources such as Wikipedia than in Wikidata and other knowledge graphs (e.g. https://en.wikipedia.org/wiki/COVID-19_pandemic_on_cruise_ships)

Applications

COVID-19 KG-driven applications

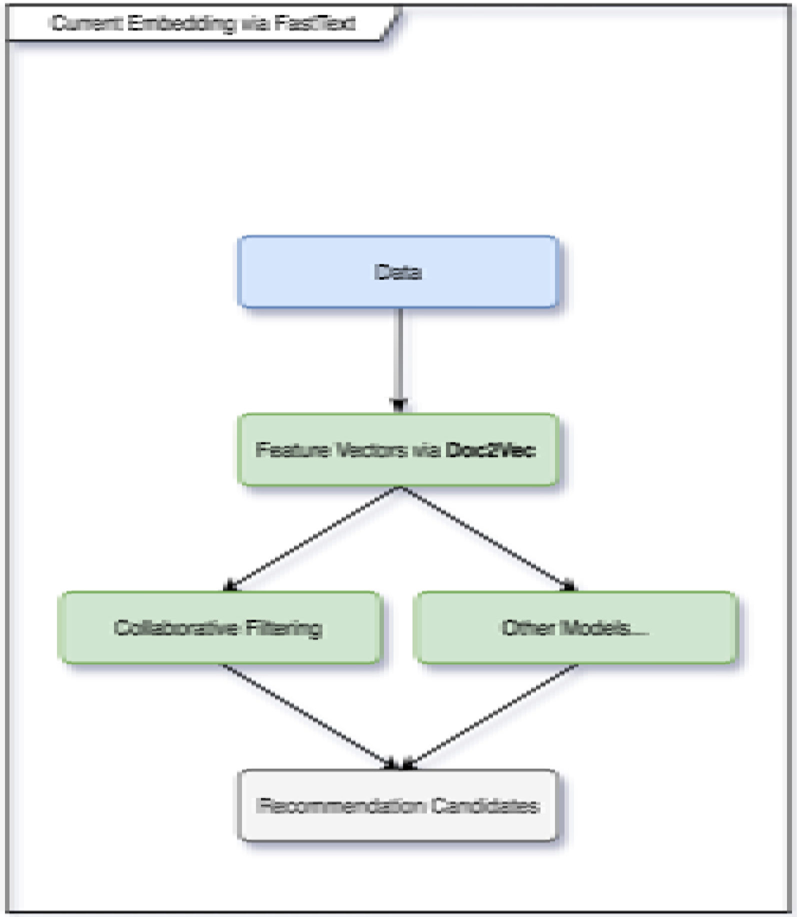
Multilingual topic modelling of social interactions

In progress



Knowledge-Based Systems

Education and Social Recommendation



Current Embedding via FastText

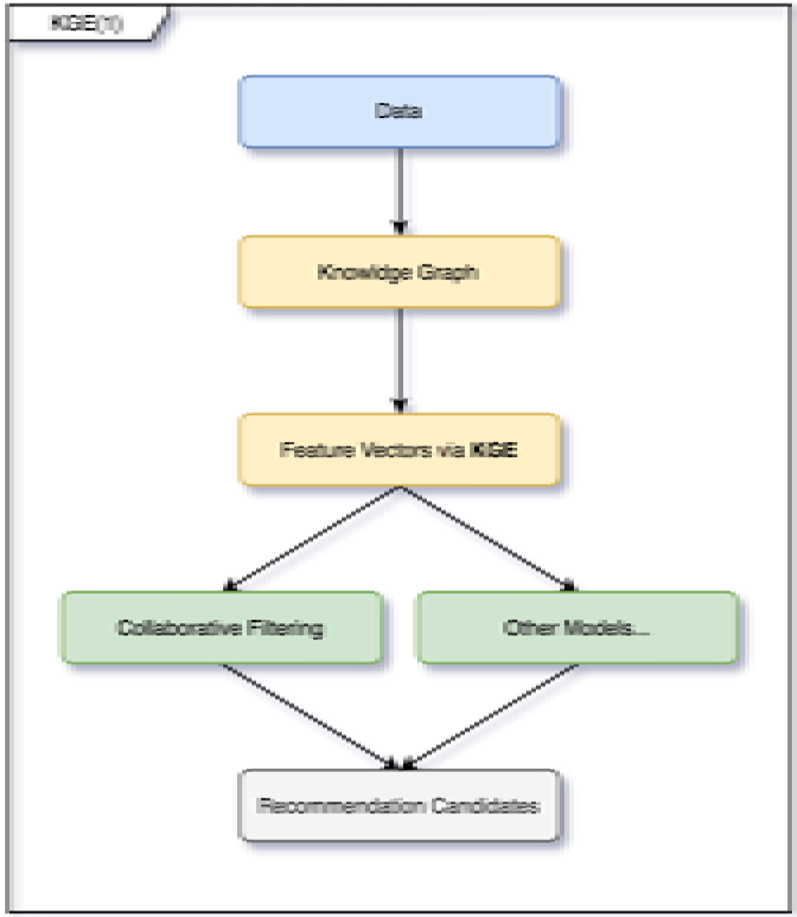
Data

Feature Vectors via **Doc2Vec**

Collaborative Filtering

Other Models...

Recommendation Candidates



KGE(1)

Data

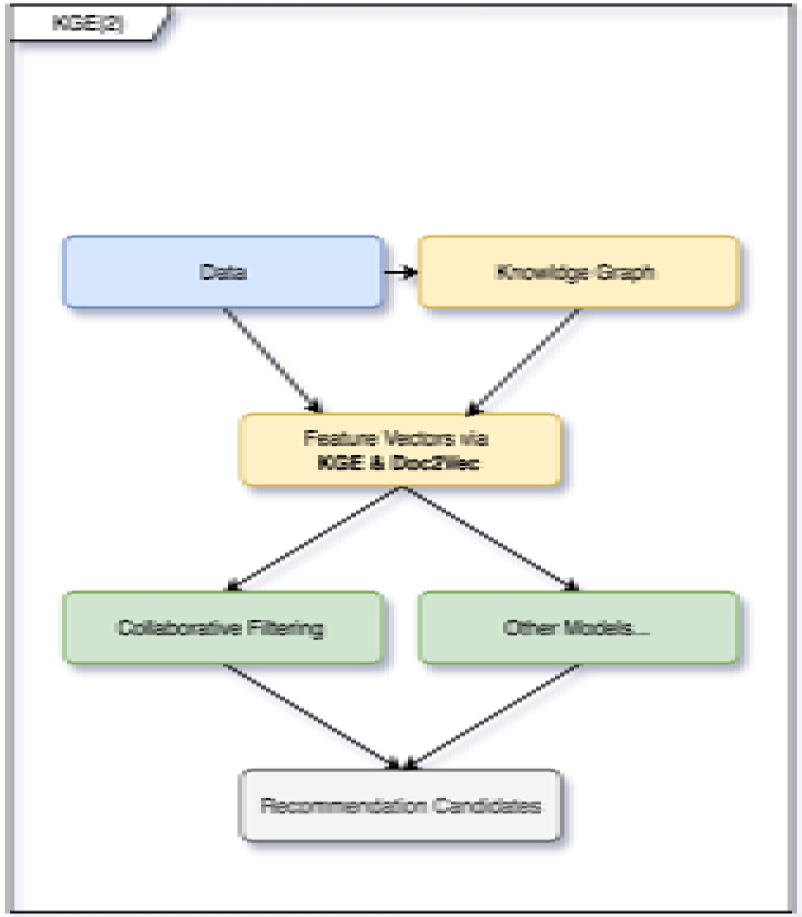
Knowldge Graph

Feature Vectors via **KGE**

Collaborative Filtering

Other Models...

Recommendation Candidates



KGE(2)

Data

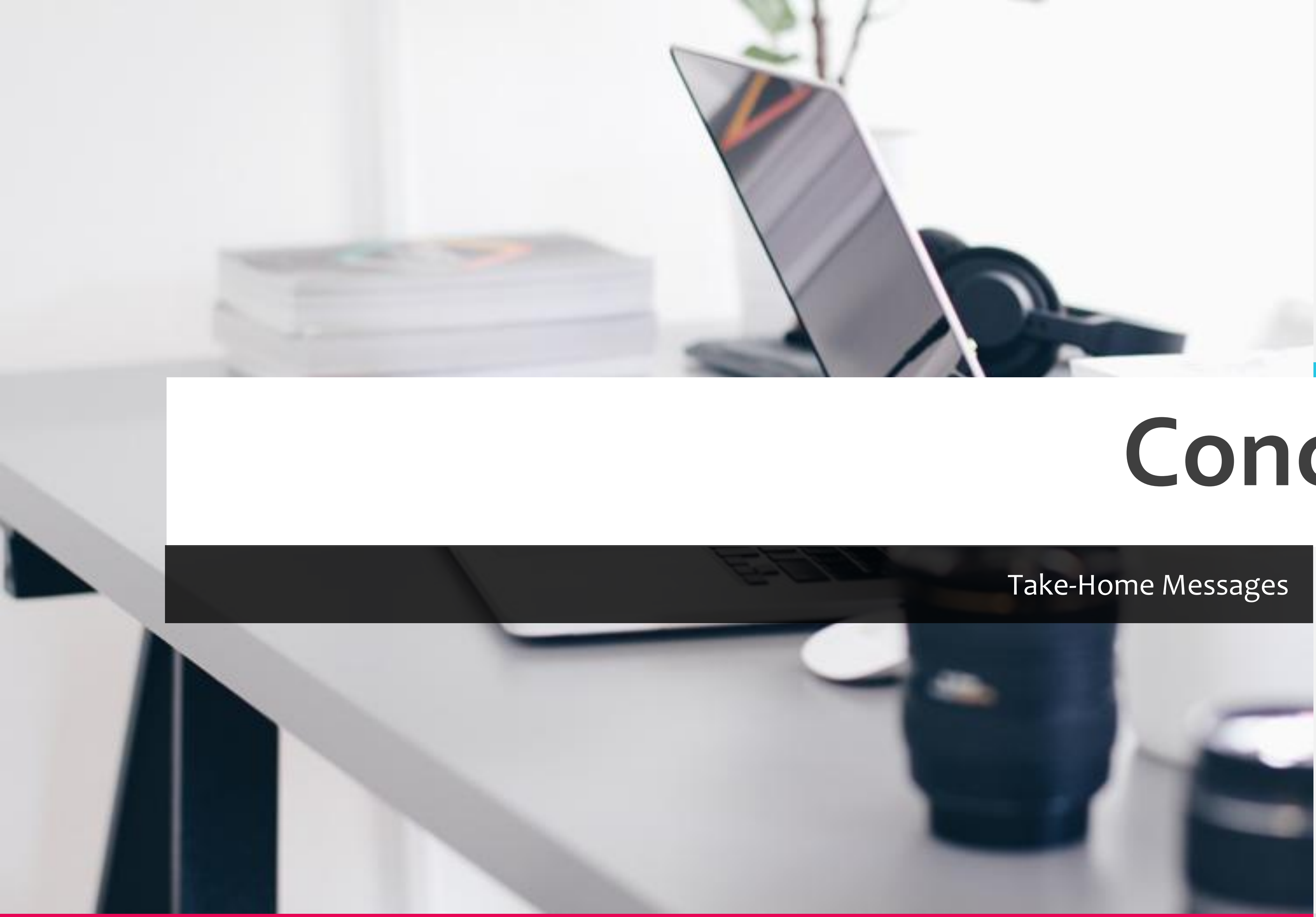
Knowldge Graph

Feature Vectors via
KGE & Doc2Vec

Collaborative Filtering

Other Models...

Recommendation Candidates



Conclusion

Take-Home Messages

Conclusion

- Wikidata and other collaborative multidisciplinary knowledge graphs can create more efficiently a semantic database for COVID-19.
- Despite their slight limitations, Collaborative multidisciplinary knowledge graphs like Wikidata can return interesting findings about COVID-19 due to the integration of COVID-19 multidisciplinary information with non-COVID-19 information.
- Due to its interesting coverage of COVID-19, Wikidata can be used for a variety of applications using semantic web tools.

To cite the work

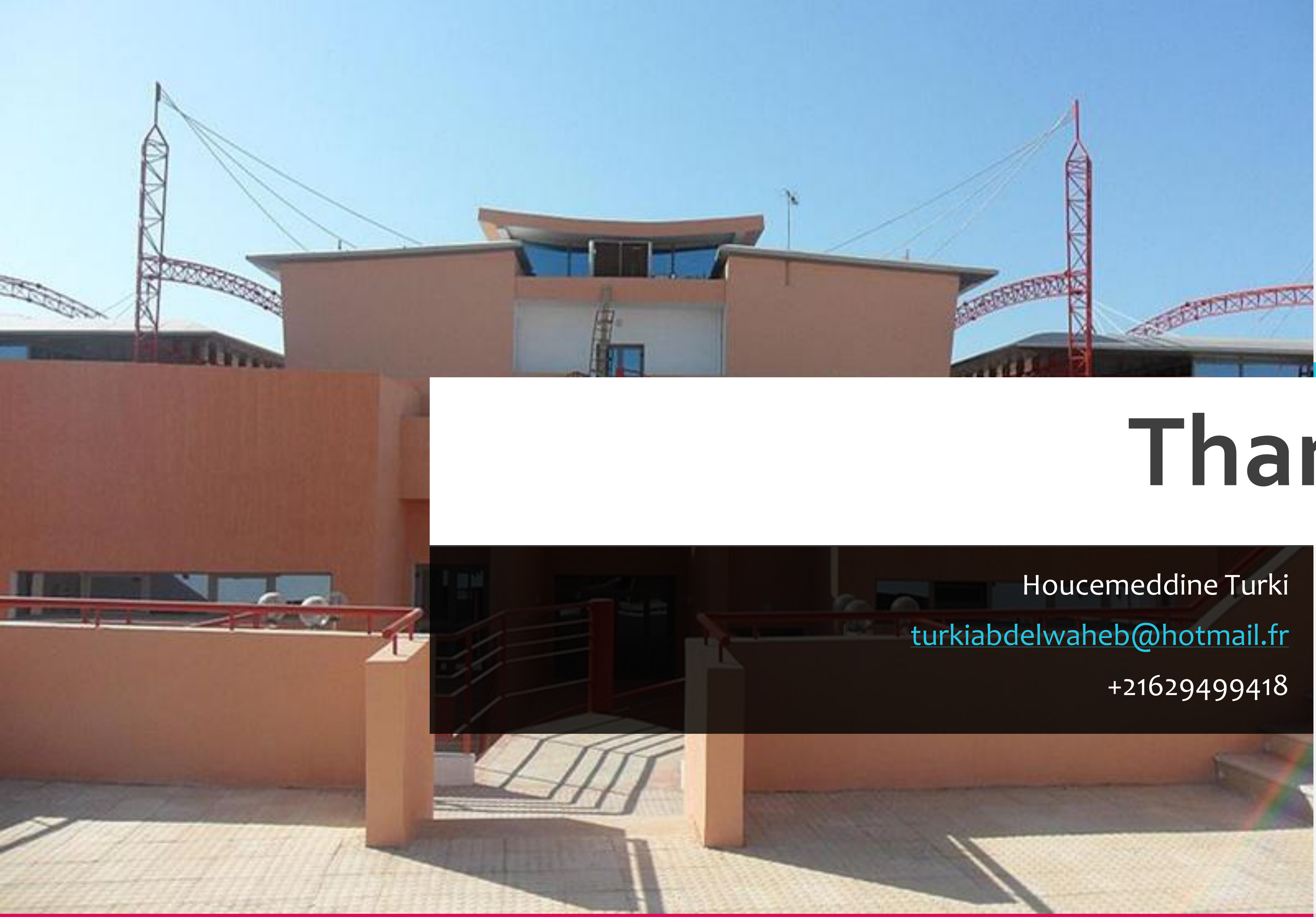
- **Main Work:**
  - Turki, H., Shafee, T., Hadj Taieb, M. A., Ben Aouicha, M., Vrandečić, D., Das, D., & Hamdi, H. (2019). Wikidata: A large-scale collaborative ontological medical database. *Journal of biomedical informatics*, *99*, 103292. doi:10.1016/j.jbi.2019.103292.
  - Turki, H., Hadj Taieb, M. A., Shafee, T., Lubiana, T., Jemielniak, D., Ben Aouicha, M., Labra Gayo, J. E., Banat, M., Das, D., & Mietchen, D. (2020). Representing COVID-19 information in collaborative knowledge graphs: a study of Wikidata. *Zenodo*. doi:10.5281/zenodo.4028482.
  - 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. (2020). Using logical constraints to validate information in collaborative knowledge graphs: a study of COVID-19 on Wikidata. *Zenodo*. doi:10.5281/zenodo.4008358.
  - Waagmeester, A., Willighagen, E. L., Su, A. I., Kutmon, M., Gayo, J. E. L., Fernández-Álvarez, D., ... & Koehorst, J. J. (2020). 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., Griffith, M., Griffith, O. L., ... & Keating, S. M. (2020). Science Forum: Wikidata as a knowledge graph for the life sciences. *ELife*, *9*, e52614. doi:10.7554/eLife.52614.
- **Applications:**
  - Xianxian, H. (2019). Knowledge Graph (KG) for Recommendation System. *Medium*. https://medium.com/@hxianxian/knowledge-graph-kg-for-recommendation-system-8fe2c6cd354.

References

- Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78-85. doi: 10.1145/2629489.
- Wimmics Research Team (2020). Covid-on-the-Web dataset. *Zenodo*. doi:10.5281/zenodo.3833752.
- Michel, F., Gandon, F., Ah-Kane, V., Bobasheva, A., Cabrio, E., Corby, O., ... & Simon, M. (2020, November). Covid-on-the-Web: Knowledge Graph and Services to Advance COVID-19 Research. In *International Semantic Web Conference*.
- CNRS (2020). Multivac Platform. *GitHub*. https://github.com/multivacplatform.
- Colavizza, G., Costas, R., Traag, V. A., Van Eck, N. J., Van Leeuwen, T., & Waltman, L. (2020). A scientometric overview of CORD-19. *BioRxiv*. doi:10.1101/2020.04.20.046144.
- Lastra-Díaz, J. J., Goikoetxea, J., Hadj Taieb, M. A., García-Serrano, A., Ben Aouicha, M., & Agirre, E. (2019). A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art. *Engineering Applications of Artificial Intelligence*, *85*, 645-665. doi:10.1016/j.engappai.2019.07.010.
- 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.
- Peng, Y., Yan, S., & Lu, Z. (2019, August). Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task* (pp. 58-65).
- Piad-Morffis, A., Estevez-Velarde, S., Estevanell-Valladares, E. L., Gutiérrez, Y., Montoyo, A., Muñoz, R., & Almeida-Cruz, Y. (2020). Knowledge Discovery in COVID-19 Research Literature. *OpenReview*. https://openreview.net/forum?id=CWfGhEFOTKU
- Wu, Y., Liu, M., Zheng, W. J., Zhao, Z., & Xu, H. (2012). Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. In *Biocomputing 2012* (pp. 422-433). doi:10.1142/9789814366496_0041.
- Piad-Morffis, A., Gutiérrez, Y., & Muñoz, R. (2019). A corpus to support ehealth knowledge discovery technologies. *Journal of biomedical informatics*, *94*, 103172. doi:10.1016/j.jbi.2019.103172.
- 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. doi:10.1016/j.jbi.2018.05.011.

Credits

- https://commons.wikimedia.org/wiki/File:Du_cot%C3%A9_du_bab_diwan.jpg
- https://creativecommons.org/licenses/by/4.0/
- http://www.webdo.tn/2018/10/16/luniversite-de-sfax-1ere-en-tunisie-801eme-dans-le-monde/
- https://commons.wikimedia.org/wiki/File:A_Rainbow_Of_Books_-_Flickr_-_Dawn_Endico.jpg
- Colavizza, G., Costas, R., Traag, V. A., Van Eck, N. J., Van Leeuwen, T., & Waltman, L. (2020). A scientometric overview of CORD-19. *BioRxiv*. doi:10.1101/2020.04.20.046144.
- https://commons.wikimedia.org/wiki/Category:COVID-19_Study_of_Wikidata
- Xianxian, H. (2019). Knowledge Graph (KG) for Recommendation System. *Medium*. https://medium.com/@hxianxian/knowledge-graph-kg-for-recommendation-system-8fe2c6cd354.
- https://commons.wikimedia.org/wiki/File:Int%C3%A9rieur_2_du_centre_de_recherche,_Technopole_de_Sfax.jpg



Thank You

Houcemeddine Turki

turkiabdelwaheb@hotmail.fr

+21629499418

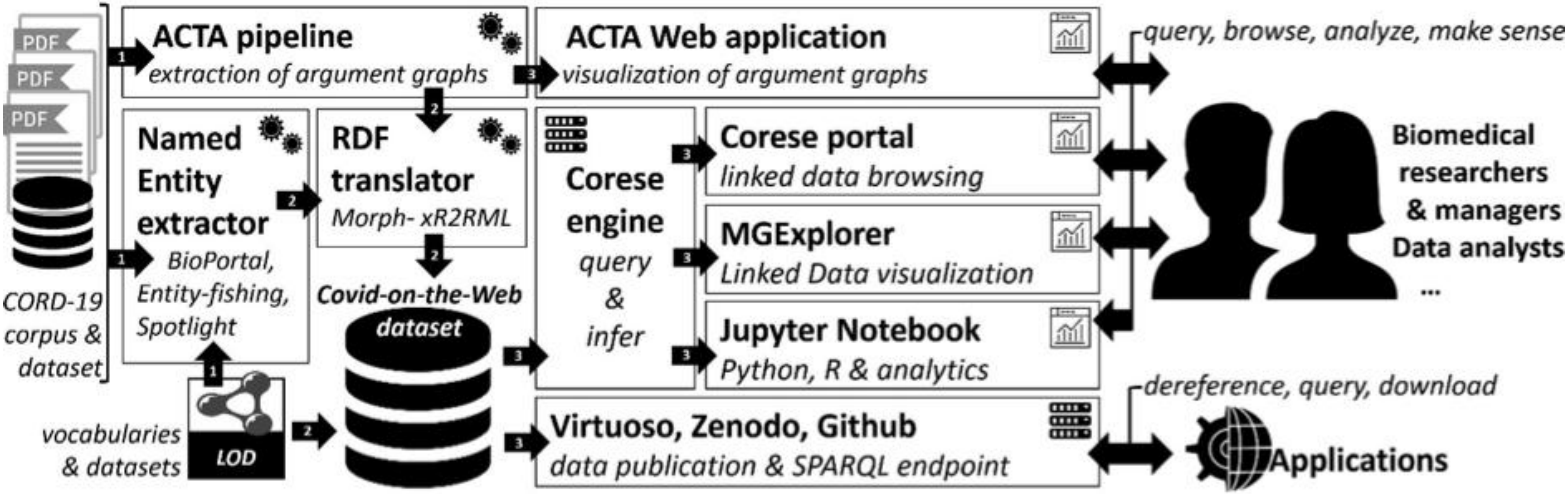
Appendices

Useful information

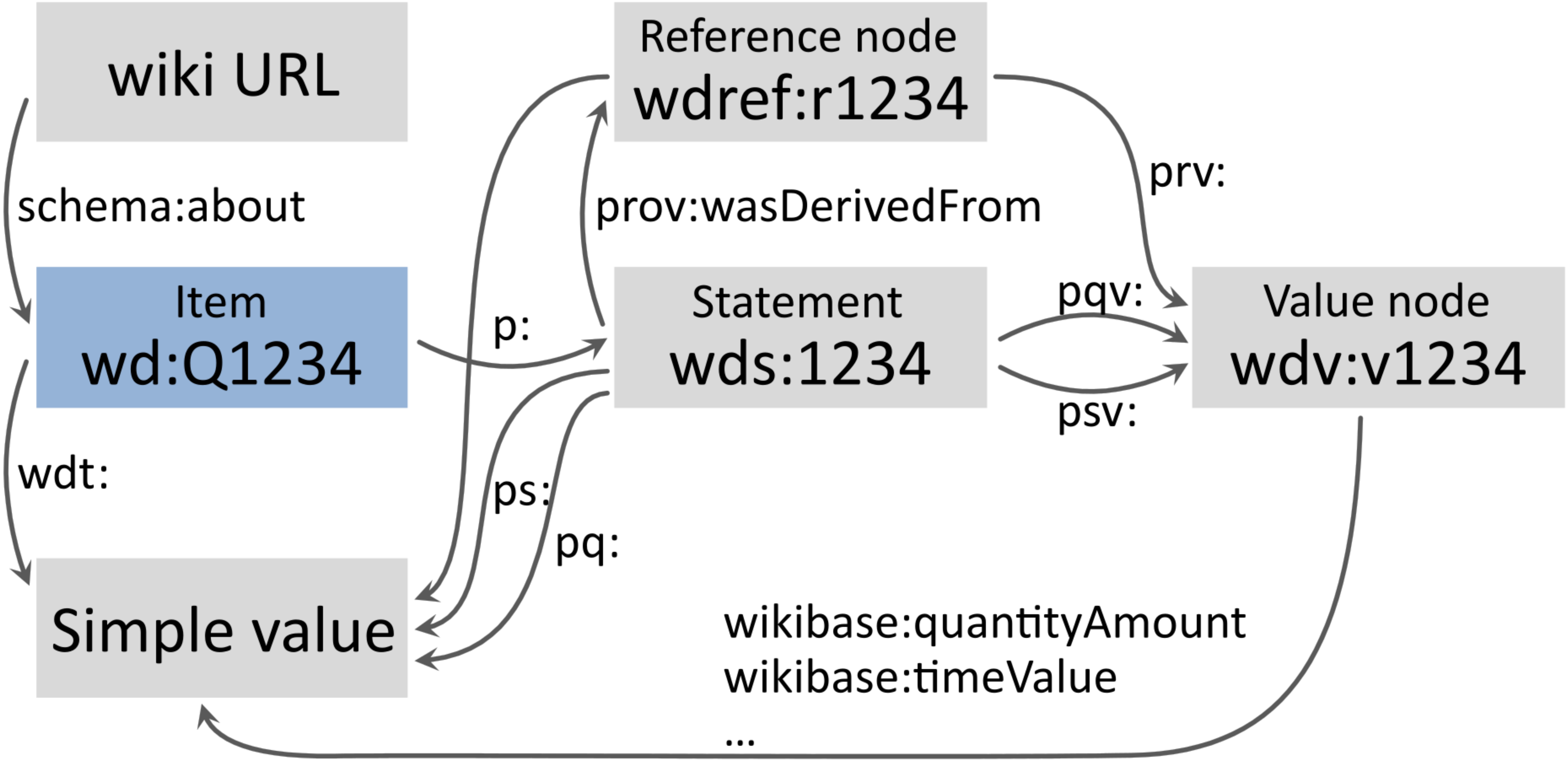
- Appendix A: Algorithm for the development of CORD-19 NEKG
- Appendix B: RDF data format for Wikidata
- Appendix C: Wikidata prefixes
- Appendix D: Useful links
- Appendix E: Hidden Tools - FM3S
- Appendix F: Hidden Tools - SNOWL Model

Appendix A: Algorithm for the development of CORD-19 NEKG

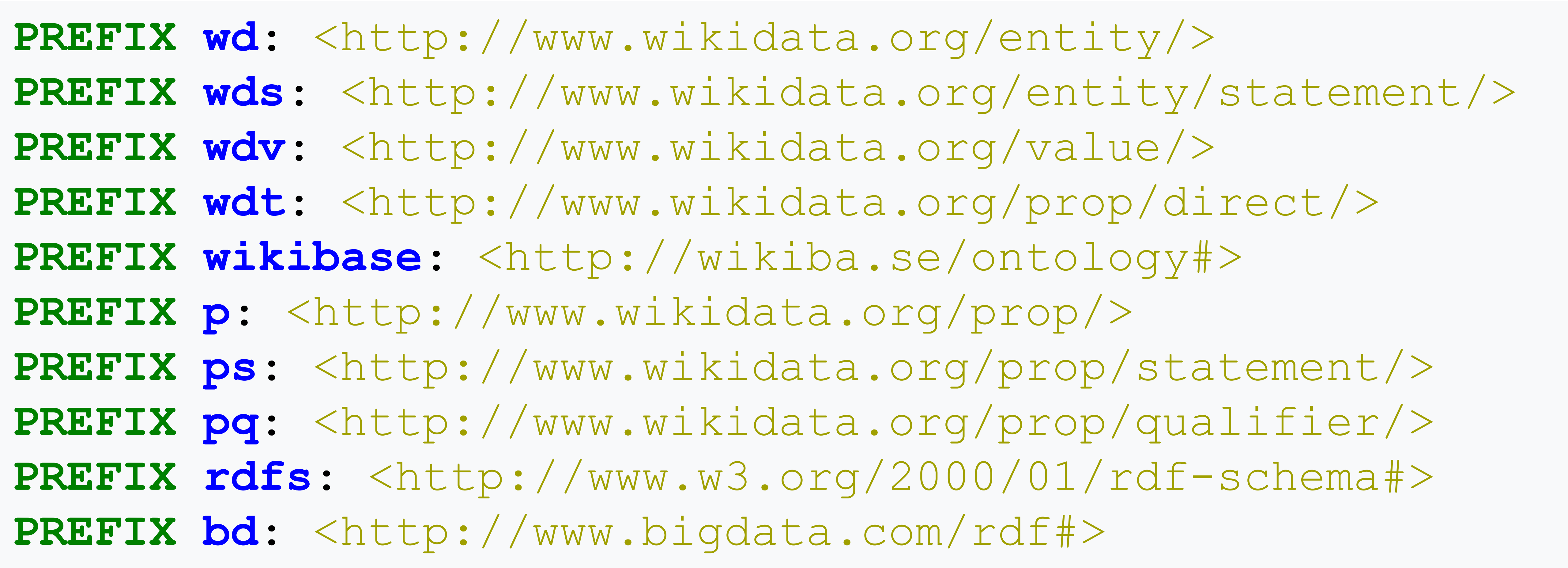
https://www.inria.fr/fr/covid-web



Appendix B: RDF data format for Wikidata



Appendix C: Wikidata prefixes



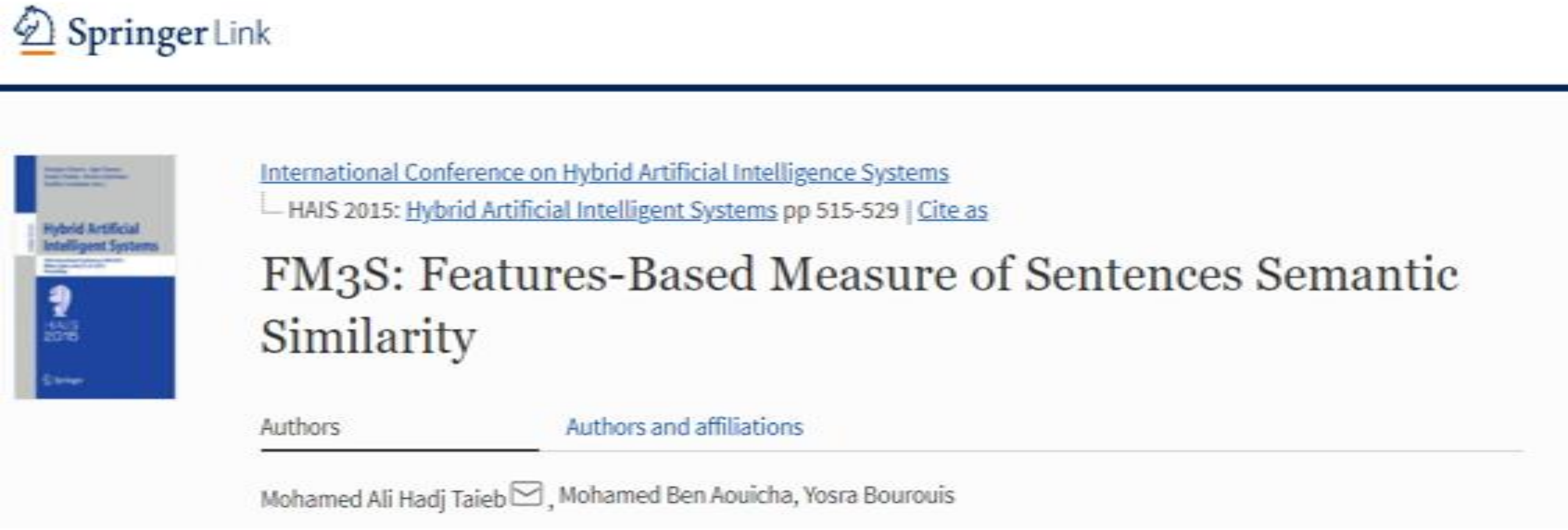
```
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wds: <http://www.wikidata.org/entity/statement/>
PREFIX wdv: <http://www.wikidata.org/value/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX ps: <http://www.wikidata.org/prop/statement/>
PREFIX pq: <http://www.wikidata.org/prop/qualifier/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX bd: <http://www.bigdata.com/rdf#>
```

Appendix D: Useful links

Page	URL
Statistics	https://www.wikidata.org/wiki/Special:Statistics
Database Reports	https://www.wikidata.org/wiki/Wikidata:Database_reports
Database download	https://www.wikidata.org/wiki/Wikidata:Database_download
User access levels	https://www.wikidata.org/wiki/Wikidata:User_access_levels
Wikidata Tour	https://www.wikidata.org/wiki/Wikidata:Tours
Wikidata Tools	https://www.wikidata.org/wiki/Wikidata:Tools
Wikidata Hub	https://hub.toolforge.org/
Scholia tool (Scholarly information)	https://scholia.toolforge.org/
SPEED tool (Epidemiological information)	https://speed.ieee.tn/
SARS-CoV-2-Queries (Genomics)	https://egonw.github.io/SARS-CoV-2-Queries/
Bot flag for RefB (Bot adding references to biomedical statements)	https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/RefB_(WikiCred)
Source code for RefB	https://github.com/Data-Engineering-and-Semantics/refb/

Appendix E: Hidden Tools - FM3S

A measure of sentence-level semantic similarity



Springer Link

International Conference on Hybrid Artificial Intelligence Systems

HAIS 2015: Hybrid Artificial Intelligent Systems pp 515-529 | Cite as

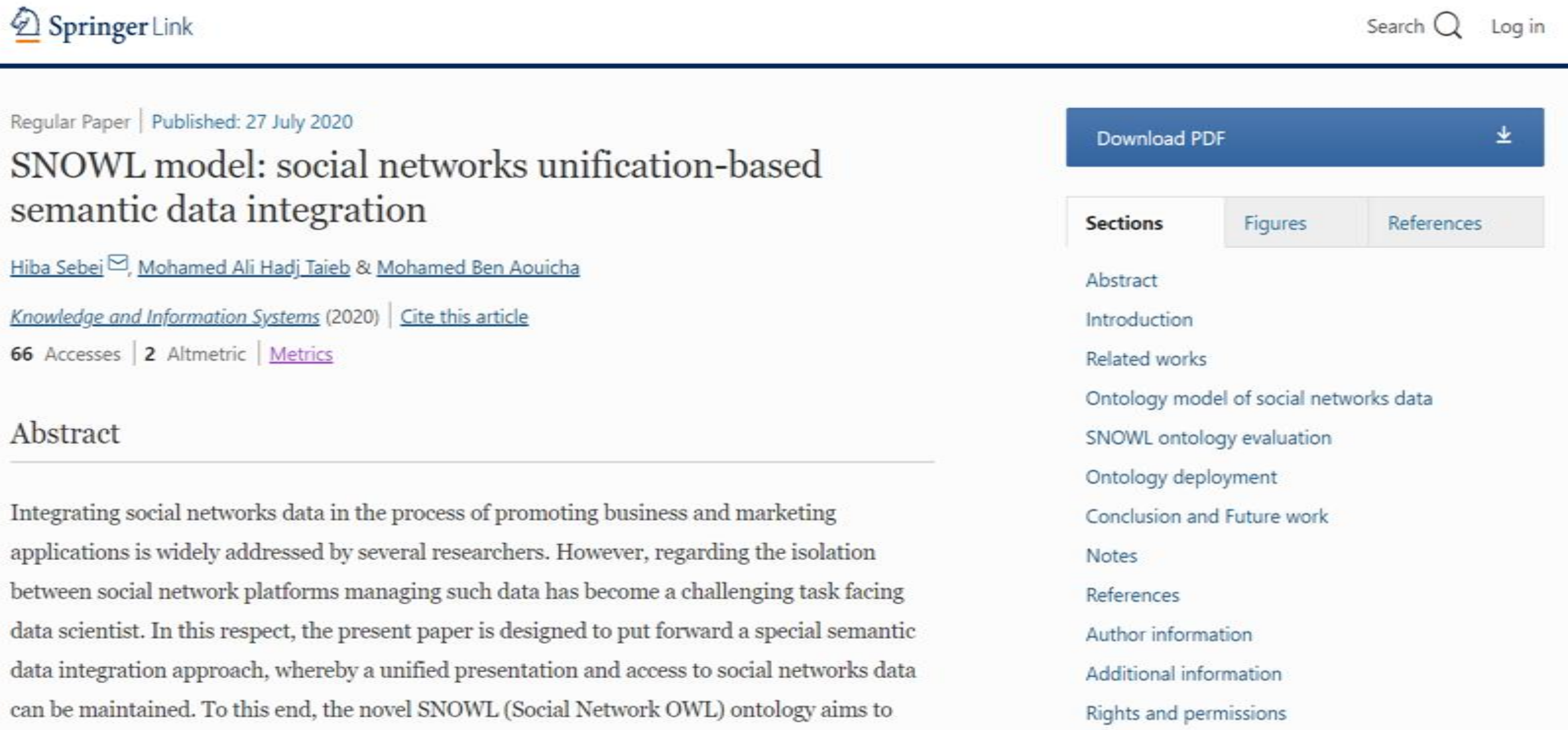
FM3S: Features-Based Measure of Sentences Semantic Similarity

Authors | Authors and affiliations

Mohamed Ali Hadj Taieb ✉, Mohamed Ben Aouicha, Yosra Bourouis

Appendix F: Hidden Tools - SNOWL Model

An ontology for the alignment between the namespaces and entity types in social media



Regular Paper | Published: 27 July 2020

SNOWL model: social networks unification-based semantic data integration

Hiba Sebei, Mohamed Ali Hadj Taieb & Mohamed Ben Aouicha

Knowledge and Information Systems (2020) | Cite this article

66 Accesses | 2 Altmetric | Metrics

Abstract

Integrating social networks data in the process of promoting business and marketing applications is widely addressed by several researchers. However, regarding the isolation between social network platforms managing such data has become a challenging task facing data scientist. In this respect, the present paper is designed to put forward a special semantic data integration approach, whereby a unified presentation and access to social networks data can be maintained. To this end, the novel SNOWL (Social Network OWL) ontology aims to