AGILE: GIScience Series, 3, 21, 2022. https://doi.org/10.5194/agile-giss-3-21-2022
Proceedings of the 25th AGILE Conference on Geographic Information Science, 2022.
Editors: E. Parseliunas, A. Mansourian, P. Partsinevelos, and J. Suziedelyte-Visockiene.
This contribution underwent peer review based on a full paper submission.

COVID-Forecast-Graph: An Open Knowledge Graph for Consolidating COVID-19 Forecasts and Economic Indicators via Place and Time

Rui Zhu [1,2], Krzysztof Janowicz[1,2,3], Gengchen Mai [1,2,4], Ling Cai [1,2], and Meilin Shi [1,2]

[1]STKO Lab, Department of Geography, University of California, Santa Barbara, USA
[2]Center for Spatial Studies, University of California, Santa Barbara, USA
[3]Department of Geography and Regional Research, University of Vienna, AT
[4]Department of Computer Science, Stanford University, USA

Correspondence: Rui Zhu (ruizhu@ucsb.edu)

Abstract. The longer the COVID-19 pandemic lasts, the more apparent it becomes that understanding its social drivers may be as important as understanding the virus itself. One such social driver is misinformation and distrust in institutions. This is particularly interesting as the scientific process is more transparent than ever before. Numerous scientific teams have published datasets that cover almost any imaginable aspects of COVID-19 during the last two years. However, consistently and efficiently integrating and making sense of these separate data "silos" to scientists, decision makers, journalists, and more importantly the general public remain a key challenge with important implications for transparency. Several types of knowledge graphs have been published to tackle this issue and to enable data crosswalks by providing rich contextual information. Interestingly, none of these graphs has focused on COVID-19 forecasts despite them acting as the underpinning for decision making. In this work we motivate the need for exposing forecasts as a knowledge graph, showcase queries that run against the graph, and geographically interlink forecasts with indicators of economic impacts.

Keywords. COVID-19, place and time, interoperability, knowledge graphs, knowledge representation

1 Introduction

A pandemic such as the Spanish flu that started in 1918 is not merely an entirely different event than the ongoing COVID-19 pandemic for medical and biological reasons, but also due to changes in social structures, the densely interlinked supply chains of our global economy, and most notably in the ways in which information is communicated. Instead of a limited number of largely authoritative data sources that report data occasionally, we are bombarded with thousands of heterogeneous sources from governments, research labs, the industry, media, and individuals, with an update frequency of minutes, not days. Naively, one may assume that this makes for more informed citizens and decision making as the transparent process in which data are reported by many of these organizations enables the comparison of reports and forecasts. Instead, we are facing two pandemics, a medical and a social one at the same time. There may be different reasons for this such as distrust in authorities, fake news, misinformation campaigns, social inequality, the type of rewards systems powering social media more broadly, and so on (Roozenbeek et al., 2020; Cuan-Baltazar et al., 2020; Tasnim et al., 2020). For example, one current work warns us that with richer countries are preparing to lift public-health interventions, widely open the economy, and report data less frequently thanks to the success of vaccination campaign, there are still billions of people that are vulnerable to the pandemic (Mathieu, 2022). Providing open, clean, and reliable data to the general public regarding the COVID-19 pandemic will be a long-term mission for the whole society.

We believe that one way to support citizens and also decision makers in the industry, NGOs, and even local governments, is to provide access to pre-integrated data that visualize relationships across multiple COVID-19 facets such as case loads and economic impacts. For instance, a key argument brought up by those that questioned lockdown at the early stage of the pandemic was the types of places that are being forced to close while others remain open. Put differently, it is difficult to put together medical, economic, governmental, and other factors in a coherent picture. One such example is the aggregation of California's Central Coast with its relatively lower population density and lower case loads with the very heav-

ily affected and very densely populated Southern California region around Los Angeles. Understandably, the communities that form the Central Coast rallied for the creation of a separate region as it would have allowed them to keep their economies open (Brune, 2020). Another example is the dozens of different forecasts with their underlying assumptions such as the introduction or easing of restrictions. To citizens and even the press, it often remains unclear where numbers that are used to justify new measures or relax old one are taken from. One early example comes from Germany. In the fall 2020, former Chancellor Angela Merkel urged the public to return to stricter measures as "coronavirus infection rate could hit 19,200 per day in Germany if the current trend continues" (Michelle, 2020). This statement was met with wide-spread criticism even from the press and was called alarmist as Germans could not imagine that numbers would climb that high. As it turned out just weeks later, the daily cases reported in Germany easily surpassed this estimate. A major concern raised by the press was a lack of transparency. It took days to find the particular forecast Chancellor Angela Merkel was referring to in her press conference. With an ever increasing number of research teams attempting to study the spread and impacts of the COVID-19 pandemic, a comparison of the models underlying these forecasts and their underlying, often hidden, assumptions, e.g., the widespread usage of vaccines, will be key to understanding the trend of the pandemic for months or even years to come. For the United States alone, there are 114 models published by about 93 different teams (as of February 2022) to predict cumulative and incident deaths, as well as the number of hospitalization[1]. Each of these models comes with own assumptions, different spatial and temporal scales, and is updated with new measures being introduced, variants being detected (e.g., Omicron BA2), and with additional vaccines becoming available. These models also range widely in their overall fatality rates, at times even by a full order of magnitude and as will be discussed below. Finally, their accuracy varies substantially over *time* and with *geographic space*, i.e., across regions.

In this work, we develop a geographically-enabled knowledge graph for COVID-19 forecasts and an ontology to semantically represent the ingested data. We enrich this graph with data collected about the various assumptions underlying forecast models and the type of modeling they use, e.g., Machine Learning versus Susceptible, Exposed, Infectious, and Recovered (SEIR) models. Next, we integrate this graph with economic indicators. The entire graph is indexed both geographically and temporally. Finally, to demonstrate the value of our work, we will showcase several exemplary queries against the graph, some with surprising results. The graph, ontology, testing queries, visualizations, as well as the lifting source code are available online[2]. To support full SPARQL queries over the graph, we made it available as a public query endpoint[3].

2 Related Work

Over the past two years, a great number of data sets has been introduced to help understand and mitigate the global COVID-19 pandemic from various aspects. Particularly in the Scientific Data community[4], more than 32 data descriptors, analysis, or comments have been published, including topics from government intervention and policy (Desvars-Larrive et al., 2020; Zheng et al., 2020; Porcher, 2020; Shiraef et al., 2021), human mobility (Zheng et al., 2020; Kang et al., 2020), biomedical studies, (Ellinger et al., 2021; Liu et al., 2021; He et al., 2020; Desai et al., 2020), and those on psychology impacts (Yamada et al., 2021; Mondino et al., 2020; Sugaya et al., 2020; Bailon et al., 2020). However, despite its importance in decision-making, few data sets are available that describe forecast of COVID-19 spread, not to mention its cross-walks with economic indicators, census data, and so on. Meanwhile, several teams have developed knowledge graphs (Hogan et al., 2020) to give access to pre-integrated data related to COVID-19, such as case loads, impacts on transportation, scientific literature, and drug repurposing (Michel et al., 2020; Wang et al., 2020; Domingo-Fernández et al., 2020). These graphs aim at establishing an ecosystem of data that involves multiple disciplines so that a single piece of information can be timely *enriched* by data from other disciplines following *FAIR* principle (i.e., findable, accessible, interoperable, and reusable). Interestingly, none of these knowledge graphs targets forecasts and their underlying assumptions despite their key roles in decision-making.

3 Methodology

In this work, we introduce the COVID-Forecast-Graph focusing on providing a holistic view of Web-available forecast models about COVID-19. While some platforms and news organizations provide a *visual* representation of many of these models, they differ substantially from our work in several regards. First, we publish the actual model outputs, not simply visuals. Second, while these models originate from a common API that makes them Web-available, we integrate data about the underlying assumptions behind these models from several resources, thereby enriching the source data. Third, we pre-integrate the models with other types of data such as reported "ground truth", credit card usage in business sectors, human mobility, and employment rate through the nexus of place and time at various scales. Based on the place and time component, we also provide means to easily navigate between regions. For instance, we showcase that models perform

[1] https://zoltardata.com/project/44
[2] https://github.com/zhurui0509/COVID-Forecast-Graph
[3] https://stko-roy.geog.ucsb.edu/covid
[4] https://www.nature.com/collections/ebaiehhfhg

best in selected states and no model outperforms all other models geographically. Finally, by modeling forecasts and related data using an ontology on top of the international SOSA/SSN standards (Janowicz et al., 2019; Haller et al., 2019) for the semantic representation of observation data, we aim to support interoperability, machine reasoning, and the prediction of new links on top of our graph. In this section, we first elaborate major data sources for COVID-Forecast-Graph, based on which design of the underlying ontology - COVID-SO - will be discussed.

3.1 Data Source

We collect data directly from the listed repositories below. Each of these repositories might have their own source of collecting the raw data (e.g., through modeling, manually, or from private companies). To be able to integrate different model forecasts, "ground truth" observations, and economic indicators, we represent them all within a sensor and observation framework (i.e., SOSA/SSN).

3.1.1 COVID-19 Forecast

The COVID Forecast Hub Team[5] establish an open-source repository for teams to upload their forecasts (Ray et al., 2020; Cramer et al., 2021), and meanwhile provide services such as interactive visualizations and ensemble models to facilitate teams and the general public to explore the data[6]. Thanks to them requiring consistent specifications for all uploaded data as well as their official use by the US Centers for Disease Control and Prevention (CDC) to inform decision makers and the public[7], we use this repository as the main seed source to build the proposed graph. This also enables us to keep the graph up-to-date.

Each forecast has multiple targets that aim at predicting a specific variable (e.g., cases, deaths, or hospitalizations) at a specific time (i.e., prediction time) for a range of spatial units (e.g., states and/or counties). The forecast is made for 1 through 20 weeks ahead of the prediction time for incident and accumulative death, 0 through 130 days ahead for incident hospitalization, and 1 through 8 weeks ahead for incident cases. In addition to a single point prediction, the variables are represented as a distribution with 23 quantiles including 19 ones from 0.05 to 0.95 with an interval of 0.05 plus 4 extra - 0.025, 0.01, 0.90, and 0.975. Our proposed ontology captures all these various pieces of information and represents them as a knowledge graph instead of a traditional plain tabular forms. More specifically, each forecast model is regarded as the virtual sensor while its prediction results, including both single point prediction and corresponding quantiles, are represented as observations in the graph.

[5] https://COVID19forecasthub.org/

[6] https://github.com/reichlab/COVID19-forecast-hub

[7] https://www.cdc.gov/coronavirus/2019-ncov/COVID-data/forecasting-us.html

3.1.2 Daily Reported Cases and Deaths

To validate forecasts made by different teams using different methods under various assumptions, we also include daily cases and deaths reported from Johns Hopkins University's Center for Systems Science and Engineering (JHU CSSE)[8] as the "ground truth" (Dong et al., 2020). In contrast to COVID-19 forecasts, the cases and deaths are represented only as point-based values rather than a quantile distribution. These data sets have a daily temporal resolution and cover spatial units including nation, state, and county. In COVID-Forecast-Graph, we regard JHU CSSE's repository as the sensor and the reported cases and deaths as observations.

3.1.3 Economic Data

COVID-19 has a substantial impact on both the global and domestic economy. This impact is expected to last for several years. In order to build connections between COVID-19 variables, such as incident cases, with the local economic status, we collect a range of economy-related observations from the public repository of Economic Tracker[9]. These data include economic indicators such as unemployment rate, credit card use, job post, human activity, and small business revenue, which are originally collected by individual private companies (Chetty et al., 2020). The Opportunity Insights team made these data available to the public in tabular forms[10]. Our work, in contrast, regards the Economic Tracker as a virtual sensor and the economic indicators of interest as observations so as to lift the data into a knowledge graph that semantically interlinks across various data sources. These observations include nations, states, as well as counties spatial units as well as either daily or weekly temporal resolution.

3.2 Ontology Design

To improve the interoperability and reusability of these separate data "silos", we design a COVID-19 Sensor and Observation ontology, COVID-SO, on top of the W3C recommended Semantic Sensor Network ontology (SSN/SOSA) and its extensions. In this section, we first discuss key concepts and their relations defined in SSN/SOSA, based on which the proposed COVID-SO will be introduced next.

3.2.1 Semantic Sensor Network Ontology and its Extensions

The Semantic Sensor Network (SSN) ontology (Haller et al., 2019) is a W3C[11] and OGC[12] recommenda-

[8] https://coronavirus.jhu.edu

[9] https://tracktherecovery.org/

[10] https://github.com/OpportunityInsights/EconomicTracker

[11] https://www.w3.org/

[12] https://www.ogc.org/

tion/standard to represent sensors, their observations, samples, associated stimulus, and so on. The Sensor, Observation, Sample, and Actuator (SOSA) ontology (Janowicz et al., 2019) is the lightweight core of SSN, in which only the core classes and their corresponding properties are specified. SOSA has been widely applied to domains such as remotely sensed images (Kostovska et al., 2020), smart city (Espinoza-Arias et al., 2019), humanitarian relief (Zhu et al., 2021b), the Internet of Things (Honti and Abonyi, 2019), and environmental intelligence (Zhu et al., 2021a). To enable modeling homogeneous collections of observations, an extension to SSN has been introduced[13]. It uses a new collection constructor to efficiently represent observations that share common characteristics, such as the feature-of-interest, phenomenon time, sensor, and so on.

3.2.2 COVID-19 Sensor and Observation Ontology

SOSA and its extension facilitate us to integrate and reuse the cross-disciplinary and heterogeneous repositories related to the COVID-19 pandemic. Concretely, we design a three-tier ontology, with the upper-level establishing the relation between SOSA and COVID-19 related core concepts; the middle-level specifying the relation among these core concepts (some are reused from SOSA and some are newly specified); and the lower-level detailing various subclasses of the core concepts described in the middle-level. In the following, we discuss details for each level.

3.2.3 Upper-level COVID-SO.

As illustrated in Figure 1, core components of COVID-SO utilize elements of SOSA (orange boxes), and the COVID-19 related concepts (red boxes) are linked to them via rdfs:subClassOf relation so as to reuse the inherited properties. While this level mainly supports interoperability and provides us with the proper concepts to describe regions (sosa:FeatureOfInterest), time (time:TemporalEntity), observable properties (sosa:ObservableProperty), etc., the specification of each individual data source is discussed in the middle-level ontology.

It is worth highlighting that the two properties - sosa:resultTime and sosa:phenomenonTime - differ in their semantics despite both describing temporal information. First, sosa:resultTime is a data type property, which links a subject to a literal data type (e.g., xsd:dateTime), while sosa:phenomenonTime is an object property, which links the subject with an object (e.g., an instance of the time:TemporalEntity class). Semantically, the sosa:resultTime is used to describe the time when a sensor makes an observation while the sosa:phenomenonTime records the time of the observed phenomenon. This difference becomes particularly essential when modeling forecasts, such as the COVID-19 forecast that will be discussed below, as (according to the specifications) the sosa:phenomenonTime is the time of the targeted prediction, which is later than the sosa:resultTime when the prediction is executed.

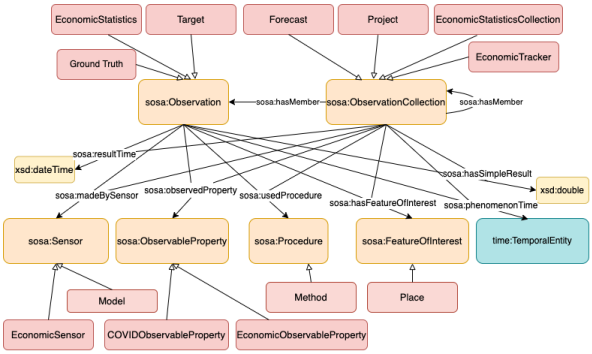

Figure 1. Upper-level ontology of COVID-SO. The prefix of covid-so for introduced classes (red boxes) and properties is removed for brevity. Hollow arrow indicates the rdfs:subClassOf property.

3.2.4 Middle-level COVID-SO.

SOSA, together with its extension, are designed with both physical and virtual sensors in mind to support simulations as well as forecasting (Lefrançois et al., 2016; Janowicz et al., 2019). As outlined above, forecasts are observations whose sosa:phenomenonTime is later in time than the corresponding sosa:resultTime. While applying SOSA specifically to COVID-19 related virtual sensors (i.e., forecast models), we have to extend it in order to fit our need of efficiently describing all aspects of the source data. More concretely, we focus on modeling the forecast, their underlying assumptions, reported "ground truth", as well as relevant economic data with the goal of interconnecting them, and potentially many other data sources as well, through the consistent use of terminologies from the COVID-SO ontology. Hence, middle-level concepts extracted from each data repository (Section 3.1) are discussed as follows.

- COVID-19 Forecast

  Figure 2 depicts the ontology to represent sensors and observations collected from the COVID-19 Forecast repository (Section 3.1.1). To illustrate it, we will use the research led by Georgia Tech University's DeepOubreak team (GT-DeepCOVID)[14] as an example. GT-DeepCOVID (covid-so:Project) made a forecast (covid-so:Forecast) on 2022-02-14 (xsd:dataTime), in which there is a specific target (covid-so:Target) that forecasts the "8 days ahead incident hospitalization" (covid-so:TargetType) for all US states. More specifically, this target predicts that the average daily new hospitalization (covid-so:COVIDObservableProperty) in the state of South

[13] https://www.w3.org/TR/vocab-ssn-ext/

[14] https://deepcovid.github.io/

Dakota (covid-so:Place) on 2022-02-22 (covid-so:TemporalEntity) will reach 38.11 (xsd:double). GT-DeepCOVID (covid-so:Project) makes such a forecast (covid-so:Forecast) using a model (covid-so:Model) built on Deep Learning method (covid-so:Method) with an assumption that the current intervention policy will remain in the place (covid-so:Assumption). The method belongs to the family of general Machine Learning techniques (covid-so:MethodType). Moreover, the applied model is the primary one (covid-so:ModelDesignation - used in case a team has multiple models) and is reported by Georgia Tech University (covid-so:Team) under the Creative Common licence: CC BY 4.0 (covid-so:License). There are multiple reported funding agents (covid-so:Organization) for this team on this specific project (e.g., the US National Science Foundation).

Other projects and their forecasts are modeled in a similar way. Nevertheless, it is worth noting that different projects might target different observable properties (e.g., daily incident cases, cumulative deaths, and daily incident hospitalization) at different spatial resolutions (e.g., county, state, and national level).

Our ontology also represents the uncertainty inherent in the forecasts since most predictive models will output a quantile distribution of the estimated variable rather than a single value (see Section 3.1.1). This is achieved by introducing a list of sub-properties of the original sosa:hasSimpleResult property in SOSA. For example, most predictions made from the COVID-19 Forecast repository are represented as either one point value or a list of 23 quantiles. Therefore, we designed 24 sub-properties: covid-so:point and covid-so:quantile-N, where N is a place holder for the value of the quantile (e.g., 0.01, 0.25, 0.4, and so on).

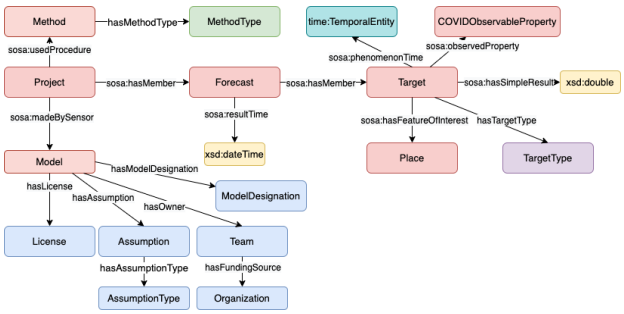

Figure 2. COVID-19 forecast in COVID-SO. Boxes in blue are designed to describe middle-level concepts about covid-so:Model; Box in green is for covid-so:Method; Box in purple is for covid-so:Target. Yellow boxes are for literal information. The prefix covid-so is removed for brevity.

- COVID-19 Reported "Ground Truth"

  The ontology for reported "ground truth" is designed in a similar fashion. It involves four basic classes: covid-so:GroundTruth, covid-so:COVIDObservableProperty, covid-so:Place, and time:TemporalEntity (see Figure 3). For instance, it can represent that the reported "ground truth" (covid-so:GroundTruth) of the cumulative number of death (covid-so:COVIDObservableProperty) in Alaska (covid-so:Place) on 2021-01-15 (time:TemporalEntity) was 987.00 (xsd:double).

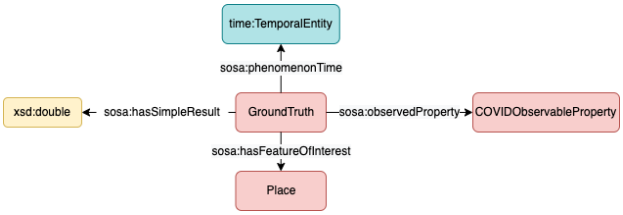

Figure 3. Reported "ground truth" in COVID-SO. The prefix covid-so is removed for brevity.

- COVID-19 Relevant Economic Data

  Likewise, we organize the semantics of economic data using middle-level components of COVID-SO. Specifically, covid-so:EconomicTracker and covid-so:EconomicStatisticsCollection are two types of sosa:ObservationCollection, and covid-so:EconomicStatistics is an instance of sosa:Observation. The reason to use observation collection here is that those statistical observations often share the same characteristics, such as being collected at the same time and same region. Using a collection can thus help reduce redundancy. Again, taking observations from one of the economic virtual sensors - Affinity[15] - as an example, we can represent its sensors and observations as: the company Affinity (covid-so:EconomicSensor) made a collection of observations (covid-so:EconomicTracker) on 2021-11-24 (xsd:dateTime), one of which is about a set of economic statistics (covid-so:EconomicStatisticsCollection) in the state of Texas (covid-so:Place) on the date of 2021-11-07 (time:TemporalEntity). In the economic statistics collection (covid-so:EconomicStatisticsCollection), one is about the seasonally adjusted credit/debit card spending in arts, entertainment, and recreation relative to January 4-31, 2020 (covid-so:EconomicObservableProperty) and its observed value is 0.402 (xsd:double).

3.2.5 Lower-level COVID-SO.

On the lower-level COVID-SO, we introduce instances that have type of the middle-level classes (represented using the property rdf:type). This level depends on specific data sources. Figure 5 depicts

[15] www.affinity.solutions

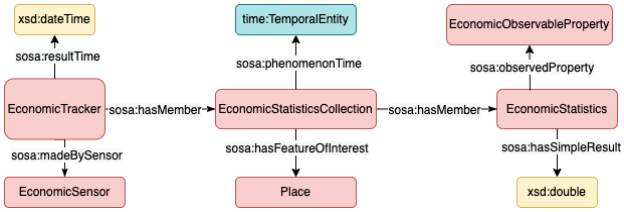

Figure 4. Relevant economic data in COVID-SO. The prefix covid-so is removed for brevity.

five examples. First of all, both covid-so:Target and covid-so:GroundTruth share the same set of covid-so:COVIDObservableProperty, which includes covid-so:incidentDeath, covid-so:cumulativeDeath, covid-so:incidentCase, and covid-so:incidentHospitalization. Secondly, data about covid-so:AssumptionType and covid-so:MethodType used in different projects are from the categorization organized by CDC[16] and we manually match the project and method names to the ones that are used in the COVID-19 Forecast repository. Regarding economic data, classes in both covid-so:EconomicSensors and covid-so:EconomicObservableProperty are obtained from the Economic Tracker Data Dictionary[17]. It is worth noting that each of the middle-level classes (e.g., covid-so:AssumptionType and covid-so:EconomicSensor) provides the entry points for including future lower-level instances and properties from new data sources.

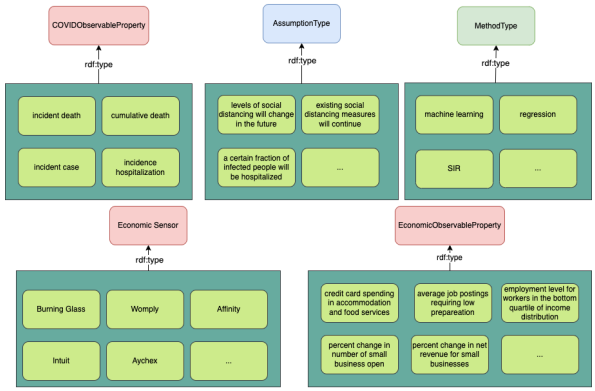

Figure 5. Ontology of lower level concepts in COVID-SO (only five examples are shown). The prefix covid-so is removed for brevity.

3.2.6 Place and Time as Nexuses for Cross-Walks

One core strength of the proposed COVID-SO is its capability to connect different data sources that are related to COVID-19. Places and time periods act as nexuses for cross-walks between and within these data sources. Figure 6 illustrates ontology fragments for both place (left) and time (right). For a place, we do not only preserve its key identifiers such as the Federal Information Processing Standards (FIPS) code and name, but also match each place to its corresponding identifier (URL) in the Wikidata knowledge graph by leveraging the *OWL:sameAs* relation. Wikidata[18] is an open knowledge graph that is built as the underlying data provider for Wikipedia. By linking to Wikidata, places in our knowledge graph are enriched with more detailed information, such as place type (e.g., county), geographic coordinate, population, mayor, ruling political party, etc. So far, the graph only includes regional administrative identifiers in the US. However, the place concept here can be extended to other identifiers, such as Point of Interest (POI) and hierarchical grid cells (Shimizu et al., 2021). Place identifiers from other countries can also be incorporated into the graph to enhance the cross-walk of COVID-19 forecast and its associated global socioeconomic indicators. Finally, places of interest in COVID-Forecast-Graph can be mainly categorized into three hierarchical scales: county-level, state-level, as well as national-level. Note that some forecasts only work on a subset of spatial scales while others observe all.

With respect to temporal information, we utilize Time Ontology in OWL – OWL-Time[19] – in order to directly take advantage of its reasoning capability. Specifically, time in our knowledge graph can be modeled as an instance of either time:Instant or time:Interval, both of which are subclass of time:TemporalEntity. In contrast to time:Interval, which represents a duration or extent of a temporal period, instances of time:Instant have no extent or duration. An interval's beginning and end points are defined as time:Instant. In COVID-Forecast-KG, if the temporal scale of an observation is daily, we regard it as a time *instant*; while if the data is an aggregation of observations across a week, the time is modeled as an *interval*. By reusing OWL-Time, we can easily retrieve and compare observations with time as an index using simple queries, which will be illustrated in the next section.

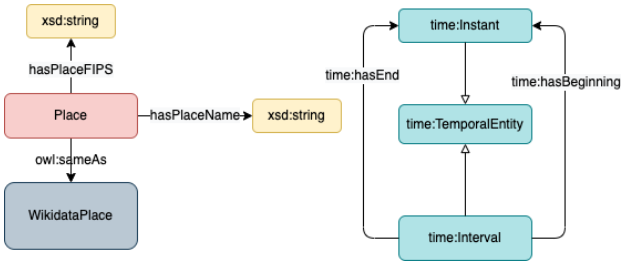

Figure 6. Ontology fragments for place (left) and time (right) in COVID-SO.The prefix covid-so is removed for brevity.

[16] https://github.com/cdcepi/COVID-19-Forecasts

[17] https://github.com/OpportunityInsights/EconomicTracker/blob/main/docs/oi_tracker_data_dictionary.md

[18] https://www.wikidata.org/wiki/Wikidata:Main_Page

[19] https://www.w3.org/TR/owl-time/

4 Result and Case Study

4.1 Statistics about COVID-Forecast-Graph

At the date of submitting this paper (February 2022), COVID-Forecast-Graph consists of $1,835$ forecasts involving 117 research projects that were made by 94 teams starting on 2021-07-01. These forecasts include predictions months into the future and the database is updated on a rolling basis. We have also included 8 economic trackers that jointly generate about $1,002,750$ economic statistics in total that are related to COVID-19. These observations are made at about $3,200$ places including states and counties in the United States in about $1,241$ temporal snapshots including dates and weeks. There are in total $220,120,470$ statements in COVID-Forecast-Graph. More detailed statistics can be found in Table 1. COVID-Forecast-Graph is updated weekly by including newly observed forecasts, reported "ground truth", and economic statistics.

4.2 Competency Questions

One strength of the presented knowledge graph is that it makes complex queries across models and datasets available at users' fingertips. In this section, we discuss multiple exemplary competency questions together with their corresponding SPARQL (a standard query language for RDF-based knowledge graphs[20]) queries. These queries result from discussions with humanitarian relief specialist at Direct Relief, an international non-governmental organization, as well as researchers in food systems and supply chain, and hence only reflect a small, but representative, fraction of potential queries. These queries are also clustered into four main groups based on their goals. Note that the part inside of the bracket (i.e., [...]) can be replaced by other instances from the same class. Plus, all discussed exemplary queries can be found at the published Github repository as well as the graph endpoint and can be readily executed.

Group 1: Description of sensors and observations

Q1: *Which projects have forecasts about [cumulative death] for [California] on [2022-03-12]? When have these forecasts been made? Which of them used [regression analysis]?*

Once a model is designed and published, it will be used to predict the variable of interest frequently. However, since these projects are developed by various teams and are maintained in different approaches, not all of them are applied for predictions at the same frequency (e.g., once a week, twice a week, or random). Hence, it is common to pick forecasts at a specific place and time and then observe the changes among these forecasts only. Once users have identified a relevant project and its models, e.g., by their past accuracy, more contextual information can be queried via the graph such as the type of method used, the funding resource, the geographic location of the team behind the project and so on (see Q2 for an example).

Q2: *Which projects implement the assumption that [local social distancing policies will be kept in place]? Which methods do these projects utilize?*

Forecast models that are developed by different research projects vary in their applied assumptions and underlying mathematical methods. Therefore, in order to understand how a model works or which types of models are more robust over time and geographic space, obtaining knowledge about both the assumptions and method types becomes imperative. A model with an assumption that the social distancing will be lifted in the next 4 weeks or that vaccines will be distributed rapidly would not work well for states that do not plan to enact such a policy or where rollout is slow. Even though many data sources have provided metadata files, they are often stored separately from the core data, lack a consistent format, and are not organized in a way ready for ingestion by downstream models or visualizations. For example, such information often resides as unstructured text in FAQs. This creates a significant barrier for end users to efficiently capture the context of using a specific model and its forecasts, putting the transparency and trustiness of decision-makings based on these data sources into questions. Our COVID-Forecast-Graph addresses this challenge by allowing end users to use just one query to answer questions such as Q2 that requires meta-level information about a predication (e.g., underlying model assumptions and used methods).

Group 2: Comparison between sensors and observations

Q3: *Find all predicted [cumulative death] in [California] on [2022-02-12], and compare them with the reported [ground truth].*

Given all the different forecasts, a decision-maker might want to first figure out which model performs best for a specific region at a given time. By doing so, all forecasts at that place and given time have to be found first, and then they have to be compared with the reported "ground truth". This process can be achieved by running a simple query on COVID-Forecast-Graph. Moreover, since most forecast models report both the average prediction and its quantile distribution (i.e., uncertainty), we also provide an example query to help users evaluate a model on whether it intends to predict a relatively short interval that includes the reported "ground truth".

[20] https://www.w3.org/TR/rdf-sparql-query/

Key class	Project	Forecast	Target	Economic Tracker	Economic Statistics Collection	Economic Statistics	Place	Temporal Entity
Number of entities	117	1,835	12,920,792	9	116,448	1,265,922	3,200	1,241
Total number of statements: 221,339,868 Total number of entities: 17,217,023 Total number of properties: 62 Total number of classes: 29								

Table 1. Statistics of COVID-Forecast-Graph as of 2022-02-18. Prefixes are removed from the class name for brevity.

Q4: *Among all the [4-week ahead forecasts] of [cumulative death] in [early January 2022 (i.e., before the Omicron peak in the US)], which model performed the best for each state across the US?*

Similar to Q3, users might want to further compare the best model across all states in the US. Since states might have employed different policies, differ in population density, mobility, international connectivity, and so on, which can be key assumption underlying a model, we hypothesize that for different models, their performances should be rather distinct across states. There are many ways to evaluate a model's forecasting capability. In this work, we showcase one possible solution using the inter-connected COVID-Forecast-Graph. Specifically, we first extract the earliest forecast date of each research project in January 2022, and then use the accuracy (e.g., absolute loss) of predicting the cumulative death in the next 4 weeks to find the "best" model for each state. Figure 7 illustrates the striking results. It becomes clear that no single model performs best and that there are geographic patterns underlying the performance of successful models. Similar analyses can be conducted at the spatial scale of counties, as well as on the temporal dimension as the performance of a model for a specific state might also change through time based on ever-changing local factors such as policy, economic status, medical resource, and so on. Please note that these differences are substantial, i.e., the best model for a state varies substantially from the second best model and so on. For instance, the best model for California is about 91 fatalities (per day) off, the second best increases to 210, while the worst model is about 7909 cases off (See Table 2).

Rank	Model	Absolute Delta (in fatalities)
1	MIT_ISOLAT-Mixtures	91.45
2	MIT_CritData-GBC	210.00
3	IHME-CurveFit	219.19
...	...	...
34	UMich-RidgeTfReg	7908.64

Table 2. Model comparison of predicting COVID-19 cumulative death per day in January 2022 for California, US.

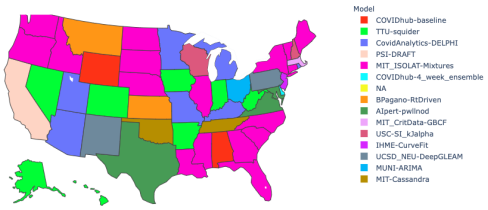

Figure 7. The best forecast model to predict the cumulative death in the next 4 weeks for each state in the US (estimated in the month of January 2022).

Q5: *In which US state does the [JHUAPL-Bucky] model perform the best (and the worst) compared with other models on [2022-02-05]? How do the results differ from forecasts for [2022-02-12] (another target forecast date)?*

For a specific model, such as the Bucky by Johns Hopkins University Applied Physics Lab[21], a decision-maker may wonder in which states (or counties) this model works best and where it falls short. To answer this question, our COVID-Forecast-Graph allows one to query and rank the performance of *JHUAPL-Bucky* in comparison to all other available models for each state at a specific target date. More concretely, Figure 8 depicts the ranking class (e.g., top 10%) of *JHUAPL-Bucky* in the ordered list of all available models' performances (i.e., absolute loss) in forecasting the cumulative death on 2022-02-05 and 2022-02-12, respectively, for each US state. For instance, we observe that *JHUAPL-Bucky* remarkably outperforms other available models in Mississippi for the two selected dates with both being ranked in the top 10% (there are about 35 available models for the target date, so top 10% indicates that the model is ranked in the top 4 of the ordered list of model performances), while the model's performance in California is consistently poor (i.e., ranked bottom 10%) for both dates. Nevada shows a different picture where *JHUAPL-Bucky's* performance positions in the bottom 30% on 2022-02-05 but increases to the top 30% on 2022-02-12. In fact, running this type of analy-

[21] https://docs.buckymodel.com/en/latest/

sis on several models, we are able to demonstrate a substantial variation of model performances across space (regions) and time, and such an analysis can be achieved readily using SPARQL query on top of COVID-Forecast-Graph.

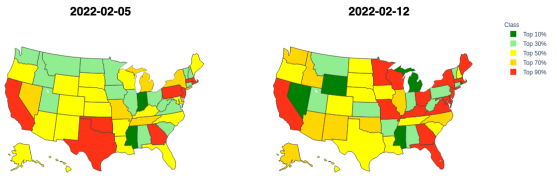

Figure 8. Performance comparison of *JHUAPL-Bucky* model across US states on 2022-02-05 and 2022-02-12.

Group 3: Integration across sensors and observations

Q6: *What is the relation between [reported incident cases] and the [time spent at retail and recreation locations] in [New York]?*

When it comes to developing policies, the spread and socio-economic impact of COVID-19 require knowledge not only about epidemiology but also from domains such as human mobility, economy, as well as local policy. Furthermore, local decision-makers and analysts do not only need the most accurate forecast model (see Q4) but also other essential information such as the employment rate, small business revenue, human activity, and credit card use in different retail sectors. Therefore, integrating all these data sources into one platform and providing efficient query capabilities are beneficial to facilitate optimal decisions. By querying reported incident cases together with the time people spent (relative to time spent in January 2020) at retail and recreation locations as an ordered time sequence for the state of New York, users can subsequently build visualizations such as time series (see Figure 9) to investigate the interaction of these two variables. For example, we can observe that the trend for visiting retail or recreation locations in New York reaches to a relatively high level thanks to the low incident cases in late 2021 while it dramatically drops due to the peak of Omicron variant in January 2022. This type of analysis can potentially help researchers investigate the impact of COVID-19 on the local economy and vice verse, which might further advance the development of new forecasting models that take into account economic indicators. It is worth emphasizing that in contrast to traditional siloed databases, integrating more repositories can be done with ease by using COVID-Forecast-Graph as it makes use of global identifiers and provides disambiguation, e.g., for places.

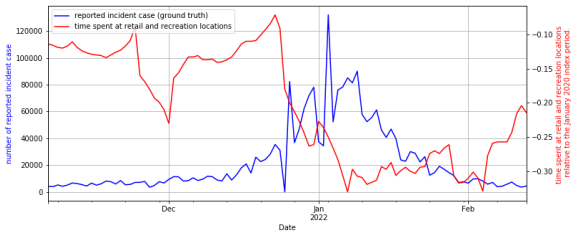

Figure 9. Time series of reported incident cases and people's time spent at retail and recreation locations (compared to January 2020) in New York.

Q7: Which model shows the largest deviation in accuracy of forecasting [cumulative death] as a function of [population density]?

In addition to economic indicators, we have also linked census related data[22] into COVID-Forecast-Graph. Consequently, one can explore the association of model performances with census-based observations (e.g., population density). For instance, Figure 10 compares the correlation between model accuracy in forecasting the number of cumulative death in 2022-02-12 and the population density (at state level) over different models. As illustrated, 34 out of 35 models hold a positive correlation between the model accuracy and population density, with the model *CU-scenario_low* achieving the strongest correlation that reaches roughly 0.40. In contrast, the model of *RobertWalraven-ESG* have slightly negative correlations with population density. This exemplary experiment indicates that for these models that have a strong correlation (either positively or negatively) between model accuracy and population density, they may consider exploring population-related effects in order to improve their performances. Furthermore, in an attempt to diagnose a specific model that shows a strong correlation, one can refer to queries as shown in Q1 and Q2 to directly retrieve the involved method types and assumptions for deeper investigations.

5 Conclusion and Discussion

In this work we introduced the COVID-SO ontology to model sensors and observations related to the COVID-19 pandemic (e.g., forecasts, reported daily cases and deaths, as well as economic statistics). In order to enable interoperability with other graphs, COVID-SO reuses the SOSA ontology and its extensions that are standardized by W3C and OGC. On top of it, we further design a middle-level and a lower-level ontology that are specific to three commonly used data sources for COVID-19: forecasts, "ground truth", and economic data. Based on COVID-SO, we create a knowledge graph - COVID-Forecast-

[22] https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html

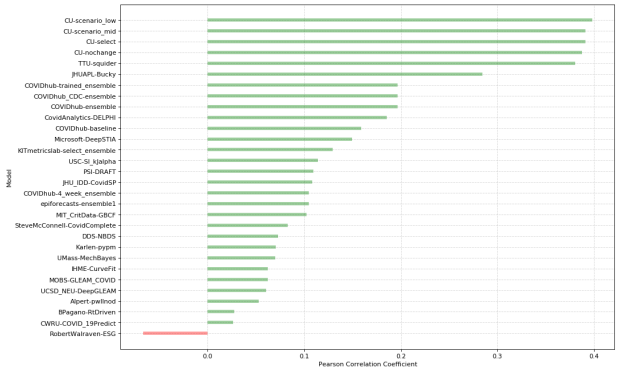

Figure 10. Comparison of the correlation between model accuracy in forecasting cumulative death on 2022-02-12 and population density across different models.

Graph - that contextualizes and enriches the aforementioned data sources so that they can be efficiently consolidated and queried. Several example queries demonstrate the strengths of COVID-Forecast-Graph.

While we included selected data sources in the current version of COVID-Forecast-Graph, the three-tier COVID-SO ontology enables further incorporation of COVID-19 data sources. This is mainly attributed to the core components of COVID-SO: *Place* and *Time*. We leverage common information about places and time periods from different data "silos" to interlink them. For example, we have integrated Wikidata into COVID-Forecast-Graph so that basic information, e.g., population density, elevation, capital cities, connectivity, temperature, water area, and so on, can be readily extracted to facilitate researchers to develop more regionally informed forecasting models. In addition, the study of SARS-CoV-2 genome (He et al., 2020; Reese et al., 2020) is another main direction of COVID-19 related research. Data sources, such as GISAID[23], have published the sequence of genes which involves information such as where it originated, when samples were collected, which lab detected them, and so on. COVID-SO potentially has the capability of integrating these data sources into COVID-Forecast-Graph so as to further empower COVID-19 related research with richer situational awareness.

Apart from the science community, stakeholders like government agencies, industry, and NGOs can contribute to as well as benefit from the proposed COVID-SO and COVID-Forecast-Graph as well. For example, the international NGO - Direct Relief[24] - has distributed over 51,000 shipments of aids globally and implemented a platform for users to track these aid recipients (Direct Relief, 2022). However, the platform and its underlying data are not pre-integrated with other data sources. Our approach of using a semantically-enriched graph enables seamless crosswalks between cases, demographics, economic indicators, and delivered relief goods.

[23] https://www.gisaid.org/

[24] https://www.directrelief.org/

6 Data and Software Availability

The COVID-SO ontology, as well as the generated knowledge graph - COVID-Forecast-Graph - are available at https://github.com/zhurui0509/COVID-Forecast-Graph. Additionally, the graph is made available at an open SPARQL endpoint (https://stko-roy.geog.ucsb.edu/covid) using GraphDB[25]. The code to generate the graph and support the application of the results of this study is available in the GitHub repository as well.

References

Bailon, C., Goicoechea, C., Banos, O., Damas, M., Pomares, H., Correa, A., Sanabria, D., and Perakakis, P.: CoVidAffect, real-time monitoring of mood variations following the COVID-19 outbreak in Spain, Scientific Data, 7, 1–10, 2020.

Brune, E.: Santa Barbara, San Luis Obispo, Ventura Counties request removal from southern California region stay home order, available at: https://publichealthsbc.org/new-page-examples/santa-barbara-san-luis-obispo-ventura-counties-request-removal-from-southern-california-region-stay-home-order/, last access: 20 February 2020, 2020.

Chetty, R., Friedman, J. N., Hendren, N., Stepner, M., et al.: The economic impacts of COVID-19: Evidence from a new public database built using private sector data, Tech. rep., national Bureau of economic research, 2020.

Cramer, E. Y., Huang, Y., Wang, Y., Ray, E. L., Cornell, M., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., House, K., Jayawardena, D., Kanji, A. H., Khandelwal, A., Le, K., Niemi, J., Stark, A., Shah, A., Wattanachit, N., Zorn, M. W., Reich, N. G., and Consortium, U. C.-. F. H.: The United States COVID-19 Forecast Hub dataset, medRxiv, https://doi.org/10.1101/2021.11.04.21265886, https://www.medrxiv.org/content/10.1101/2021.11.04.21265886v1, 2021.

Cuan-Baltazar, J. Y., Muñoz-Perez, M. J., Robledo-Vega, C., Pérez-Zepeda, M. F., and Soto-Vega, E.: Misinformation of COVID-19 on the internet: infodemiology study, JMIR public health and surveillance, 6, e18 444, 2020.

Desai, S., Baghal, A., Wongsurawat, T., Jenjaroenpun, P., Powell, T., Al-Shukri, S., Gates, K., Farmer, P., Rutherford, M., Blake, G., et al.: Chest imaging representing a COVID-19 positive rural US population, Scientific data, 7, 1–6, 2020.

Desvars-Larrive, A., Dervic, E., Haug, N., Niederkrotenthaler, T., Chen, J., Di Natale, A., Lasser, J., Gliga, D. S., Roux, A., Chakraborty, A., et al.: A structured open dataset of government interventions in response to COVID-19, medRxiv, 2020.

Direct Relief: COVID-19 Relief, available at: https://www.directrelief.org/emergency/coronavirus-outbreak/, last access: 20 February 2022, 2022.

Domingo-Fernández, D., Baksi, S., Schultz, B., Gadiya, Y., Karki, R., Raschka, T., Ebeling, C., Hofmann-Apitius, M., and Kodamullil, A. T.: COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology, Bioinformatics, p. btaa834, 2020.

[25] https://www.ontotext.com/products/graphdb/

Dong, E., Du, H., and Gardner, L.: An interactive web-based dashboard to track COVID-19 in real time, The Lancet infectious diseases, 20, 533–534, 2020.

Ellinger, B., Bojkova, D., Zaliani, A., Cinatl, J., Claussen, C., Westhaus, S., Keminer, O., Reinshagen, J., Kuzikov, M., Wolf, M., et al.: A SARS-CoV-2 cytopathicity dataset generated by high-content screening of a large drug repurposing collection, Scientific Data, 8, 1–10, 2021.

Espinoza-Arias, P., Poveda-Villalón, M., García-Castro, R., and Corcho, O.: Ontological representation of smart city data: From devices to cities, Applied Sciences, 9, 32, 2019.

Haller, A., Janowicz, K., Cox, S. J., Lefrançois, M., Taylor, K., Le Phuoc, D., Lieberman, J., García-Castro, R., Atkinson, R., and Stadler, C.: The modular SSN ontology: A joint W3C and OGC standard specifying the semantics of sensors, observations, sampling, and actuation, Semantic Web, 10, 9–32, 2019.

He, Y., Yu, H., Ong, E., Wang, Y., Liu, Y., Huffman, A., Huang, H.-h., Beverley, J., Hur, J., Yang, X., et al.: CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis, Scientific Data, 7, 1–5, 2020.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Gayo, J. E. L., Kirrane, S., Neumaier, S., Polleres, A., et al.: Knowledge graphs, Preprint at arXiv:2003.02320, 2020.

Honti, G. M. and Abonyi, J.: A review of semantic sensor technologies in internet of things architectures, Complexity, 2019, 2019.

Janowicz, K., Haller, A., Cox, S. J., Le Phuoc, D., and Lefrançois, M.: SOSA: A lightweight ontology for sensors, observations, samples, and actuators, Journal of Web Semantics, 56, 1–10, 2019.

Kang, Y., Gao, S., Liang, Y., Li, M., Rao, J., and Kruse, J.: Multiscale dynamic human mobility flow dataset in the US during the COVID-19 epidemic, Scientific data, 7, 1–13, 2020.

Kostovska, A., Džeroski, S., and Panov, P.: Semantic Description of Data Mining Datasets: An Ontology-Based Annotation Schema, in: International Conference on Discovery Science, pp. 140–155, Springer, 2020.

Lefrançois, M., Kalaoja, J., Ghariani, T., and Zimmermann, A.: The SEAS Knowledge Model, Deliverable 2.2, ITEA2 12004 Smart Energy Aware Systems, p. 76, 2016.

Liu, Y., Hur, J., Chan, W. K., Wang, Z., Xie, J., Sun, D., Handelman, S., Sexton, J., Yu, H., and He, Y.: Ontological modeling and analysis of experimentally or clinically verified drugs against coronavirus infection, Scientific data, 8, 1–12, 2021.

Mathieu, E.: Commit to transparent COVID data until the WHO declares the pandemic is over, Nature, 602, 549, 2022.

Michel, F., Gandon, F., Ah-Kane, V., Bobasheva, A., Cabrio, E., Corby, O., Gazzotti, R., Giboin, A., Marro, S., Mayer, T., et al.: Covid-on-the-Web: Knowledge graph and services to advance COVID-19 research, in: International Semantic Web Conference, pp. 294–310, Springer, 2020.

Michelle, A.: Merkel says German coronavirus infections could hit 19,200 a day, available at: https://www.reuters.com/article/us-health-coronavirus-germany-infections/merkel-says-german-coronavirus-infections-could-hit-19200-a-day-source-idUSKBN26J1FP, last access: 20 February 2022, 2020.

Mondino, E., Di Baldassarre, G., Mård, J., Ridolfi, E., and Rusca, M.: Public perceptions of multiple risks during the COVID-19 pandemic in Italy and Sweden, Scientific data, 7, 1–7, 2020.

Porcher, S.: Response2covid19, a dataset of governments' responses to COVID-19 all around the world, Scientific Data, 7, 1–9, 2020.

Ray, E. L., Wattanachit, N., Niemi, J., Kanji, A. H., House, K., Cramer, E. Y., Bracher, J., Zheng, A., Yamana, T. K., Xiong, X., et al.: Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the us, MedRXiv, 2020.

Reese, J. T., Unni, D., Callahan, T. J., Cappelletti, L., Ravanmehr, V., Carbon, S., Shefchek, K. A., Good, B. M., Balhoff, J. P., Fontana, T., et al.: KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response, Patterns, 2, 100 155, 2020.

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., Van Der Bles, A. M., and Van Der Linden, S.: Susceptibility to misinformation about COVID-19 around the world, Royal Society open science, 7, 201 199, 2020.

Shimizu, C., Zhu, R., Mai, G., Fisher, C., Cai, L., Schildhauer, M., Janowicz, K., Hitzler, P., Zhou, L., and Stephen, S.: A Pattern for Features on a Hierarchical Spatial Grid, in: The 10th International Joint Conference on Knowledge Graphs, pp. 108–114, 2021.

Shiraef, M. A., Hirst, C., Weiss, M. A., Naseer, S., Lazar, N., Beling, E., Straight, E., Feddern, L., Taylor, N. R., Jackson, C., et al.: COVID Border Accountability Project, a hand-coded global database of border closures introduced during 2020, Scientific data, 8, 1–11, 2021.

Sugaya, N., Yamamoto, T., Suzuki, N., and Uchiumi, C.: A real-time survey on the psychological impact of mild lockdown for COVID-19 in the Japanese population, Scientific Data, 7, 1–6, 2020.

Tasnim, S., Hossain, M. M., and Mazumder, H.: Impact of rumors and misinformation on COVID-19 in social media, Journal of preventive medicine and public health, 53, 171–174, 2020.

Wang, Q., Li, M., Wang, X., Parulian, N., Han, G., Ma, J., Tu, J., Lin, Y., Zhang, H., Liu, W., et al.: COVID-19 literature knowledge graph construction and drug repurposing report generation, Preprint at arXiv:2007.00576, 2020.

Yamada, Y., Ćepulić, D.-B., Coll-Martín, T., Debove, S., Gautreau, G., Han, H., Rasmussen, J., Tran, T. P., Travaglino, G. A., and Lieberoth, A.: COVIDiSTRESS Global Survey dataset on psychological and behavioural consequences of the COVID-19 outbreak, Scientific data, 8, 1–23, 2021.

Zheng, Q., Jones, F. K., Leavitt, S. V., Ung, L., Labrique, A. B., Peters, D. H., Lee, E. C., and Azman, A. S.: HIT-COVID, a global database tracking public health interventions to COVID-19, Scientific data, 7, 1–8, 2020.

Zhu, R., Ambrose, S., Zhou, L., Shimizu, C., Cai, L., Mai, G., Janowicz, K., Hitzler, P., and Schildhauer, M.: Environmental Observations in Knowledge Graphs, in: 2nd Workshop on Data and research objects management for Linked Open Science, pp. 1–11, 2021a.

Zhu, R., Cai, L., Mai, G., Shimizu, C., Fisher, C. K., Janowicz, K., Lopez-Carr, A., Schroeder, A., Schildhauer, M., Tian, Y., et al.: Providing Humanitarian Relief Support through Knowl-

edge Graphs, in: Proceedings of the 11th on Knowledge Capture Conference, pp. 285–288, 2021b.