International Journal of Information Management Data Insights 5 (2025) 100325

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei

CovKG: A Covid-19 Knowledge Graph for enabling multidimensional analytics on Covid-19 epidemiological data considering spatiotemporal, environmental, health, and socioeconomic aspects

Rudra Pratap Deb Nath [a,*], S.M. Shafkat Raihan [a], Tonmoy Chandro Das [a], Torben Bach Pedersen [b], Debasish Ghose [c]

[a] Big Data, Information and Knowledge Engineering Lab, Department of Computer Science and Engineering, University of Chittagong, Chattogram 4331, Bangladesh
[b] Department of Computer Science, Aalborg University, Aalborg 9220, Denmark
[c] School of Economics, Innovation, and Technology, Kristiania University College, Bergen, 0153, State Three, Norway

ARTICLE INFO

Keywords:
Knowledge Graph
Covid-19
Multidimensional analysis
FAIR principles
Linked data
Semantic technology

ABSTRACT

The Covid-19 pandemic is influenced by many environmental, health, and socioeconomic aspects such as air pollution, comorbidity, occupation, etc. To better manage future pandemics, decision-makers need comprehensive data on Covid-19 mortality and morbidity. Most Covid-19 data sources focus on spatiotemporal aspects, and existing research often overlook the combined impact of multiple interconnected factors. This study introduces a Covid-19 Knowledge Graph (CovKG) derived from 20 data sources, enabling multidimensional analysis of epidemiological data, including time, location, temperature, comorbidity, occupation, and others. CovKG is modeled using RDF, connected to 10,951 external resources, and semantically enriched with Data Cube (QB) and QB for OLAP (QB4OLAP) vocabularies to adhere to the FAIR principles and ensure OLAP compatibility. Finally, we perform a qualitative and comparative evaluation and extract statistical insights across multiple dimensions of Covid-19 epidemiology. When assessed, CovKG answers 100% of competency queries, outperforming other data stores that only answer 39%. CovKG and its analytical interface are available at https://bike-csecu.com/datasets/CovKG/.

1. Introduction

The Covid-19 pandemic has affected nearly every country in the world. It is a contagious disease caused by the SARS-COV-2 virus, which exhibits a high mutation and transmission rate. Since the onset of the pandemic, numerous studies, datasets, and systems have been proposed and implemented worldwide to monitor the disease's epidemiology. Epidemiology is concerned with understanding the incidence, distribution, causes, and potential control of diseases in populations. In the case of Covid-19, this can be achieved by observing attributes such as daily confirmed and death cases, recoveries, critical cases, and more. The epidemiology of Covid-19 is influenced by various factors. While most systems and studies conducted on this issue primarily focus on the spatiotemporal aspect, it is important to note that environmental, health, and socioeconomic factors also play a significant role in influencing the spread of Covid-19.

Environmental factors include variables such as air pollution, humidity, temperature, precipitation, and wind speed, which can contribute to virus incubation and influence individual health. Health-related aspects involve factors like comorbidity and vaccine hesitancy. Comorbidity, which refers to the presence of another underlying disease, can impact the immune system and consequently affect the likelihood of contracting Covid-19. The socioeconomic aspects encompass elements such as occupation, ethnicity, urbanization, and so forth. For instance, occupations that entail leaving home, such as medical professions and law enforcement, carry a higher risk of exposure to the disease. Research focusing on these different aspects has often treated them in isolation rather than considering their integrated effects. Furthermore, the hierarchical structure of these factors is frequently overlooked. For example, most data repository systems designed to monitor the spatiotemporal spread of Covid-19 only track information at the country level and do not provide details at finer sub-national levels.

In this study, we utilize Business Intelligence (BI) (Jensen, Pedersen, & Thomsen, 2010) technologies to construct a robust data framework that empowers users to comprehensively address the previously mentioned problem. This approach sets itself apart from prior studies that focused solely on individual aspects and neglected to consider the hierarchical structure of the factors involved. BI encompasses a collection

* Corresponding author.
E-mail address: rudra@cu.ac.bd (R.P. Deb Nath).

https://doi.org/10.1016/j.jjimei.2025.100325

Available online 27 February 2025

of disciplines and technological tools that provide intelligent support to decision-makers within organizations, enabling them to make efficient decisions regarding their business processes (Negash, 2004). Therefore, global organizations like the World Health Organization (WHO) can utilize BI on Covid-19 epidemiological data, including confirmed cases and deaths, to analyze and mitigate the impact of Covid-19 and its transmission.

To achieve this, the data is procured from disparate sources and integrated into a data warehouse through an Extract-Transform-Load (ETL) workflow. In the ETL workflow, data are collected from disparate sources, transformed into an agreed upon format, and loaded in a data store that allows analysis (Deb Nath, Hose, Pedersen, Romero, & Bhattacharjee, 2020; Nath, Hose, Pedersen, & Romero, 2017). Furthermore, to facilitate data analysis from multiple perspectives, the data warehouse is designed in accordance with the Multidimensional (MD) model. This model offers an easily understandable framework in which data are organized within an n-dimensional space, commonly referred to as a data cube. This space is composed of dimensions (representing the cube's axes) and facts (representing the cells within the cube) (Nath, 2020). Dimensions are ordered into hierarchies (composed of a number of levels) to explore and (dis)aggregate fact measures (i.e., numeric data) at various levels of detail (Kimball & Ross, 2011). For example, the `geographyHierarchy` hierarchy ($Admin2 \rightarrow Admin1 \rightarrow Country \rightarrow Continent$) of the `Geography` dimension allows users to (dis)aggregate the number of deaths at various administrative levels of detail.

We model the Covid-19 epidemiological data using fact constellation of data cuboids (Etcheverry, Gomez, & Vaisman, 2015) as it helps managing and modeling in a sustainable manner. It also enables Online Analytical Processing (OLAP) functionality (Chaudhuri & Dayal, 1997). OLAP functionality provides quick and accurate results. It consists of a number of operations, such as roll up (where data is aggregated to a coarser granularity), drill down (where data is dis-aggregated to a finer granularity), drill out (where data is spread out along multiple cells), drill across (where data in two cubes is merged through one or more shared dimensions), slice (where the value of one dimension level is fixed and the analysis is done along the others), dice (where the value of one or more dimension levels is fixed to one or more levels and the analysis is done along the others) etc. (Inmon, 2005; Jensen et al., 2010).

To contextualize and enable semantic integration on the data, the data warehouse is implemented as a Knowledge Graph (KG). To do this, the data cuboids are represented using the Resource Description Framework (RDF) (McBride, 2004) model. RDF is the W3C standard web data model designed for flexible data interchange on the Web. The resulting KG will truly prove to be beneficial if it is published using the principles of Findability, Accessibility, Interoperability, and Reusability (FAIR) (*FAIR Principles - GO FAIR*, 2024). A practical approach to achieve this is by publishing the data as Linked Data (Yu, 2011), which connects the KG to the Linked Open Data (LOD) cloud (*The Linked Open Data Cloud*, 2024), a network of more than 1300 interlinked datasets. To support this integration, RDF, RDF Schema (RDFS) (McBride, 2004), and Web Ontology Language (OWL) (McGuinness et al., 2004) vocabularies are used to define and enforce various constraints on the data, such as structural consistency, domain and range restrictions, class and property hierarchies, and other logical relationships. Furthermore, to annotate the data with MD semantics, the Data Cube (QB) (Tennison, Cyganiak, & Reynolds, 2012) and QB for OLAP (QB4OLAP) (Etcheverry & Vaisman, 2012) vocabularies are used. Representing the data in RDF format also enables retrieval and exploration of the data using the RDF query language, SPARQL. In summary, the unique contributions of this study are as follows.

- Producing a Covid-19 KG (CovKG) with MD semantics from diverse sources. This involves collecting data from 20 different sources, defining a target model (the schema of KG) with MD semantics by analyzing the source data, and integrating this data into the KG in a comprehensive and sustainable manner according to the semantics encoded in the target model. CovKG is also linked with other KGs available in the LOD cloud.
- Providing querying capability to identify patterns and statistics through an interactive OLAP interface. This interface allows users to construct complex OLAP queries in SPARQL seamlessly, using intuitive GUI components without requiring prior SPARQL knowledge.
- Conducting qualitative and comparative evaluation using a set of competency questions and drawing statistical insights on the multiple dimensions of Covid-19 epidemiology.

The remainder of the paper is organized as follows. Section 2 defines various terms that are frequently referred throughout the study. Section 3 discusses the previous related work. Section 4 describes the datasets and methods used in the study to model CovKG. Section 5 describes the CovKG's generation process. Section 6 describes the features of CovKG. Section 7 describes the experimental evaluation. Finally, Section 8 provides the concluding remarks and suggestions for future work.

2. Preliminaries

In this section, we introduce the relevant terms that appear frequently throughout the paper.

2.1. Knowledge Graph (KG)

A KG is a semantic graph that manifests as interlinked network of real-world entities and visualizes the relationship between them. The KG comprises two elements: Terminology Box (TBox) and Assertion Box (ABox). The TBox defines the domain schema, while the ABox represents instances (Baader, 2003). Formally, the TBox is defined as a 3-tuple: $TBox = (C, P, A^O)$, where C, P, and A^O represent the sets of concepts, properties, and terminological axioms. A concept is the blueprint of a group of instances sharing common properties. Properties establish relationships between instances of concepts or link instances of a concept to literals. Terminological axioms describe concepts, properties, and the interconnections and restrictions among them within the domain. The ABox assertions must conform to the definitions set by the TBox. In our context, the schema and instances of source datasets are called source TBoxes and ABoxes respectively. The TBox of the KG is formed by integrating and modeling the source data is the target TBox. It can have one or multiple target ABoxes. We refer to the KG consisting of the target TBox and ABoxes as the Covid-19 KG (CovKG).

In this paper, we use RDF (McBride, 2004) as the representation model for the KG. In RDF, real world entities are uniquely represented using internationalized resource identifiers (IRIs), and the description of the entities is expressed in the form of RDF triples which are three-part statements containing a subject, a predicate, and an object. For instance, in Fig. 1, the subject `cdw:SpatioTemporalDataset` has an object `admin1:dk` which is an instance of the `cdw:Admin1` class. The relation between them is expressed through the predicate `cdw:hasAdm1`. To express richer constraints on the KG, formal languages like RDFS and OWL[1] are used in combination with RDF. For example, Fig. 1 shows that in CovKG, `cdw:Admin1` is annotated as an `owl:Class`, and `admin1:dk` is externally linked to Geonames KG using the `owl:sameAs` property. Given our emphasis on MD modeling, data needs to be annotated with MD semantics at both the schematic and instance levels. For this purpose, we employ the QB4OLAP vocabulary.

[1] We use following constructs: `rdf:type`, `rdfs:subClassOf`, `rdfs:subPropertyOf`, `rdfs:domain`, `rdfs:range`, `rdfs:label`, `rdfs:comment`, `owl:Class`, `owl:ObjectProperty`, `owl:DatatypeProperty`, `owl:equivalentClass`, `owl:sameAs`.

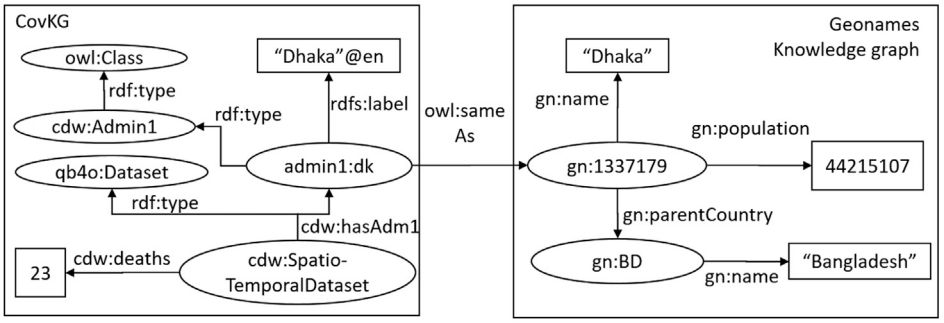


Fig. 1. A visual example of knowledge graphs. Here, `cdw:`, `admin1:`, `qb4o:`, `gn:` `rdf:`, `owl:` represent their respective namespaces.

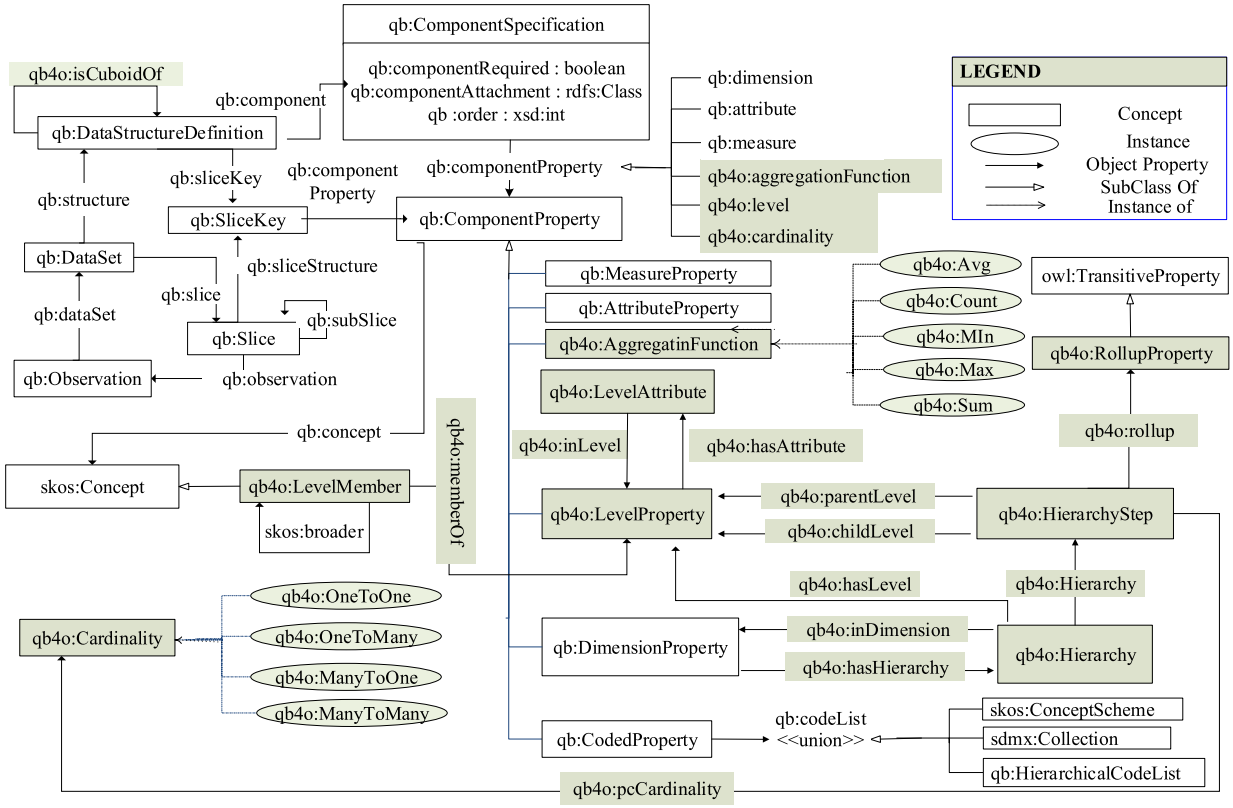


Fig. 2. The QB4OLAP vocabulary.
Source: Reproduced from Etcheverry et al. (2015).

2.2. QB4OLAP

QB4OLAP is the extension of the RDF data cube vocabulary QB (Tennison et al., 2012), which is the W3C standard for publishing statistical data as RDF. Despite being specialized for data cubes, QB does not facilitate OLAP queries to be conducted on the RDF data cubes. Thus QB4OLAP was designed as a vocabulary to represent OLAP cubes in RDF. It enables the implementation of OLAP operators as SPARQL queries directly on the RDF representation. In Fig. 2, the schematic diagram of the QB4OLAP vocabulary is given.

In the figure, the prefix `qb:` represents the terms defined in the QB vocabulary, QB4OLAP terms are denoted with the prefix `qb4o:` and displayed with a gray background. RDF classes and properties are represented by capitalized and noncapitalized terms. An arrow from class A to class B, with the label `rel` points out that `rel` is an RDF property of domain A and range B. White triangle arrows represent subclass, or subproperty relationships whereas black arrows represent `rdf:type` relationships (instances). It can be seen that the vocabulary presents various constructs to represent the dimension (`qb:DimensionProperty`), its levels (`qb4o:LevelProperty`), level members (`qb4o:LevelMember`), level attributes (`qb4o:LeveAttributes`), level's rollup properties (`qb4o:RollupProperty`), dimension hierarchies (`qb4o: Hierarchy`), and hierarchy steps (`qb4o:HierarchyStep`). In QB4OLAP, the `qb:Dataset` class is used to define an observation dataset. The structure of this dataset is outlined using `qb:DataStructureDefinition`. This structure can take the form of either a cube, if specified in dimensions and measures, or a cuboid, if outlined in levels of dimensions and measures. In Section 5, we illustrate the utilization of various QB4OLAP components for annotating CovKG with MD semantics, both in the TBox (Listing 1) and ABox (Listing 4 and 5).

3. Related work

In this section, we conduct a comprehensive investigation of prior relevant research related to the current study's topic. Through the analysis of prior studies, datasets, and systems focused on Covid-19, we categorize them into two specific groups:

1. Relevant research papers within the domain of our interest.
2. Prominent public data repositories dedicated to reporting on Covid-19.

Table 1 presents a summary of the comparative analysis of research papers and prominent data repositories regarding various features associated with Covid-19 data. The table lists the sources of previous research or repositories, indicating their involvement with confirmed and death cases, the core technologies utilized, usage of KGs, adherence to FAIR principles, compatibility with external datasets, provision of query interfaces or dashboards, availability of downloadable data, ability to conduct visual data analysis, covered aspects, and consideration of multiple dimensions. Core technologies serve as indicators of whether: (1) their processes involve either discovery techniques or surveys, (2) they employ data mining or pattern mining to uncover hidden knowledge, (3) they utilize data warehouse/OLAP technology for descriptive analysis, (4) RDF technology is used for semantic annotation, and (5) Natural Language Processing (NLP) is employed for processing scientific open data.

The table reveals that the majority of the research papers focus on analyzing confirmed cases, with only (Agapito et al., 2020) considering both death and confirmed counts. Most of these studies collect data from secondary sources and utilize data or pattern mining techniques to unveil hidden insights. However, Agapito et al. (2020) and Leung, Chen, Hoi, et al. (2020) employ data warehousing technology to enable OLAP-like analysis. Sakor et al. (2023) retrieves data from scientific literature using NLP techniques and apply KG methods to analyze drug-drug interactions. Similarly, Turki et al. (2022) also utilizes KGs, covering various aspects, although they do not enable MD analysis. Both Turki et al. (2022) and Sakor et al. (2023) adhere to FAIR principles and provide downloadable data. While most of the research papers offer query interfaces, Chen et al. (2020) and Leung, Chen, Shang and Deng (2020) do not. However, none of the studies provide a dashboard to facilitate end-users for data analysis.

Since the beginning of the Covid-19 pandemic, various government and non-government organizations worldwide have made considerable efforts to make real-time Covid-19 data available to the public. These steps have provided researchers with abundant data to conduct essential research necessary to tackle this global catastrophe. Moreover, they have made the data publicly available through online platforms, allowing users to search, view, analyze, and download data related to Covid-19. Articles (Centers for Disease Control and Prevention, 2023; COVID, DGHS, 2021; *Data - COVID-19 - Eurostat*, 2023; *WHO Coronavirus (COVID-19) Dashboard*, 2023; World Bank, 2023; Worldometer, 2020) are some of the most prominent data repositories dedicated to monitoring Covid-19 epidemiology. Some of them focus on global Covid-19 scenario, such as Worldometer (Worldometer, 2020), while there are those which target a specific region, such as Bangladesh Dynamic Dashboard for Covid-19 (COVID, DGHS, 2021). Note that the study in World Bank (2023) records number of Covid-19 waves instead of death and confirmed counts.

The studies and repositories mentioned are undoubtedly valuable, providing prolific data for analyzing Covid-19. However, as outlined in Table 1 they are not devoid of shortcomings. For instance, studies (Chen et al., 2020; Leung, Chen, Hoi, et al., 2020; Leung, Chen, Shang & Deng, 2020; Shang et al., 2020) utilized individual-level data. Although individual-level data offers a large number of influencing parameters, such parameters often have a significant amount of missing data due to undisclosed personal information. In the case of modeling occupation data, they did not employ any recognized occupation classification model. In contrast, our study aggregated population-level data—specifically, the number of confirmed cases and deaths—rather than individual cases. Such data are widely available and have the least amount of missing values. We also included occupation as one of the dimensions for analyzing Covid-19 and followed the ISCO-08 system to model it, which is an internationally renowned occupation classification system (*Classifying the Standard Occupational Classification 2020 (SOC 2020) to the International Standard Classification of Occupations (ISCO-08) - Office for National Statistics*, 2022).

In Duda, Pasichnyk, Kunanets, Antonii, and Matsiuk (2020), a theoretical framework for modeling Covid-19 data was presented, but its implementation was not realized. The authors in Turki et al. (2022) proposed an existing ontology using Wikidata to serve as a knowledge base for Covid-19. However, it is general-purpose ontology where Covid-19 information is just one facet. Our study designed and implemented a novel ontology (TBox of CovKG) and CovKG, dedicated and specialized for Covid-19 data modeling.

The data warehouse modeled in Agapito et al. (2020), named COVID-WAREHOUSE, specifically focuses on data from Italy. Our ontology covers twenty-two countries, with Italy being one of them. It also analyzes Covid-19 from thirteen perspectives, including weather and air pollution aspects. Furthermore, authors in Sakor et al. (2023) contributed to the study of relation between comorbidities and Covid-19, but they did not consider structured data. Our ontology is based on a general-purpose, structured dataset of confirmed cases and deaths, providing the flexibility to conduct interdisciplinary research with ease. The prominent data stores share the common limitation that data is not available in a semantic format. Our CovKG is a semantic data warehouse created using the RDF model. Users can link its components to external KGs. This makes data sharing and new knowledge discovery relatively easier.

In summary, our study addresses these research gaps through the application of MD semantic data integration, forming, and analyzing CovKG from multiple perspectives at fine granularities, and aims to integrate it as part of linked open data.

4. Methodology and knowledge graph modeling

The technical methodology choices in the CovKG study are based on integrating multidimensional epidemiological data into a KG for enhanced analytical capabilities (Raihan, Nath, & Das, 2023). The entire methodology to generate CovKG by modeling Covid-19 epidemiological data in an MD format is illustrated in Fig. 3. Initially, the data is collected from various data sources, which are obtained from the Web. Most of the sources present their data in CSV, XLS(X), and JSON formats. After collecting the raw data, we semantically design the TBox of CovKG (data warehouse) with MD semantics using a demand-driven approach. Then, the CovKG is populated using an ETL pipeline. In the ETL pipeline, the relevant data is extracted from the sources, transformed semantically according to the semantics encoded in the target schema, and finally, the transformed data is loaded into the data warehouse in the form of a FAIR-compliant KG. Using an OLAP interface, the OLAP operability of the CovKG is assessed. Finally, using SPARQL queries, qualitative assessment and statistical analysis are performed. Through the integration of data warehousing, OLAP functionality, and semantic modeling, this approach facilitates interoperability, scalability, and in-depth analysis of Covid-19 data across spatiotemporal, environmental, health, and socioeconomic aspects.

In this section, we outline data sources and design the target TBox for CovKG, and the next section outlines the generation of CovKG.

Table 1
Overview of related work.

Category	Reference	Death info	Confirmed cases info	Core Technologies					KG?	FAIR?	Compatible with external dataset?	Query interface/ Dash board	Enable visual Data Analysis?	Data Downloadable?	Covered aspects	Covered multiple dimensions?
				Data collection	Pattern/ data mining	DW / OLAP	RDF	NLP								
Research Papers	Shang, Leung, Chen, and Pazdor (2020)	×	✓	DD	×	×	×	×	×	×	×	✓	×	×	SP	×
	Chen, Leung, Shang, and Wen (2020)	×	✓	DD	×	×	×	×	×	×	×	×	×	×	T	×
	Leung, Chen, Shang and Deng (2020)	×	✓	DD	✓	×	×	×	×	×	×	×	×	×	SE	×
	Leung, Chen, Hoi, Shang, and Cuzzocrea (2020)	×	✓	DD	✓	✓	×	×	×	×	×	✓	×	×	SE	×
	Agapito, Zucco, and Cannataro (2020)	✓	✓	DD	×	✓	×	×	×	×	×	✓	×	×	E	✓
	Sakor et al. (2023)	×	✓	DD	✓	×	✓	✓	✓	✓	✓	✓	×	✓	H	×
	Turki et al. (2022)	×	✓	DD	×	×	✓	×	✓	✓	✓	✓	×	✓	SP, SE, H	×
Prominent Data Repositories	Worldometer (2020)	✓	✓	S	×	×	×	×	×	×	×	✓	✓	×	SP	×
	WHO Coronavirus (COVID-19) Dashboard (2023)	✓	✓	S	×	×	×	×	×	×	×	✓	✓	✓	SP	×
	World Bank (2023)	×	×	S	×	×	×	×	×	×	×	✓	✓	✓	SP	×
	COVID, DGHS (2021)	✓	✓	S	×	✓	×	×	×	×	×	✓	✓	✓	SP	×
	Data - COVID-19 - Eurostat (2023)	✓	×	S	×	×	×	×	×	×	×	✓	✓	✓	SE, SP	×
	Centers for Disease Control and Prevention (2023)	✓	×	S	×	×	×	×	×	×	×	✓	✓	×	SP	×
CovKG		✓	✓	DD	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	SP, T,E, H, SE	✓

[DD-Data Discovery, S-Survey][†] [SP-Spatial, T-Temporal, E-Environment, H-Heath, SE-Socioeconomic].

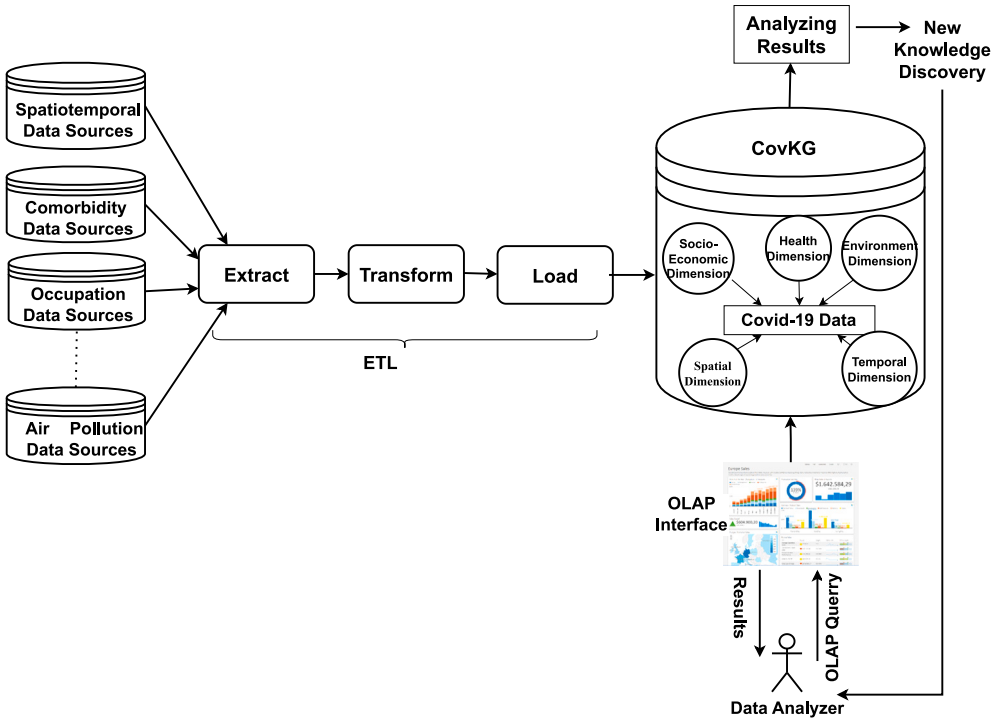


Fig. 3. Incorporation of the ETL workflow to form CovKG from disparate data sources, followed by conducting analysis on CovKG.

4.1. Description of data sources

We use different searching techniques (Baumgartner, Gatterbauer, & Gottlob, 2018; Chakrabarti et al., 2016; Dalvi, Kumar, & Soliman, 2011) for discovering Covid-19 related datasets (Roh, Heo, & Whang, 2019). During the search process, we highlight the aspects of data and selectively focus on datasets that capture information from multiple dimensions. For instance, 'Covid-19 daily province level confirmed case dataset by occupation' is an example of a self-explanatory query. We extract the data either directly from the source sites or by utilizing Application Programming Interfaces (APIs). The datasets are available as either CSV, spreadsheets, or JSON files. The data mainly consist of Covid-19 confirmed cases and death counts. The data are collected with a focus on spatiotemporal, environmental, health, and socioeconomic aspects. An overview of the data sources is provided in Table 2.

Spatiotemporal dataset: Spatiotemporal data are collected from 11 sources spanning 18 countries. Additionally, data from Turkey and Romania, obtained for the occupation dataset, are also used. Daily data for Austria, Italy, Lithuania, Poland, Slovakia, and Sweden are available at the state, province, county, voivodeship, region, and country levels, respectively. Data for Croatia, Denmark, and Finland are available at the country, municipality, and region levels, respectively. Data for Greece, Ireland, and the Netherlands are available at the prefecture, county, and municipality levels. Indian data are available at the state and union levels, whereas South African data are at the provincial level. Finally, country-level data for the USA are collected.

Weather dataset: Historic weather data for various geographic locations specified by latitude and longitude are available. The collected data contains information such as temperature (degrees Celsius), humidity (%), precipitation (millimeters), and wind speed (kilometers per hour).

Air pollution dataset: We collect the daily state-level air pollution data for South Africa and India. The pollutants considered are ground-level ozone (O3), particulates ($PM_{2.5}$ and PM_{10}), sulfur dioxide (SO_2), carbon monoxide (CO), and nitrogen dioxide (NO_2). These pollutants are commonly measured in monitoring air pollution, as per (Gadekar, 2022).

Vaccine hesitancy dataset: The questionnaire data were obtained from the Imperial College London YouGov Covid-19 Behaviour Tracker Data Hub (Jones, 2020), which records and categorizes participants' hesitant behaviors toward vaccination. Although the survey seeks to broadly represent the general public in 17 countries, our analysis concentrates on 10 countries with participants who had previously contracted Covid-19. Table 3 summarizes the distribution of survey participants by country, gender, and population representation.

Comorbidity dataset: Monthly death counts at the state level, grouped by comorbidity in the USA are available.

Ethnicity dataset: The USA death counts of various races at the state level on a monthly basis are available.

Place of death dataset: Monthly death counts at the state level based on place of death in the USA are available.

Occupation dataset: Occupation datasets are collected following the ISCO-08 standard for occupation classification. Data on COVID-19-related deaths across various occupations in England and Wales have also been gathered. Additionally, data from Romania and Turkey are available at the provincial level.

Urbanicity Dataset: Urbanicity indicates how rural or urban a region is. The urbanicity dataset is constructed from the spatiotemporal data collected for the USA by mapping counties to levels of urbanicity using the urban–rural classification scheme for the USA counties (*Data Access - Urban Rural Classification Scheme for Counties*, 2023).

4.1.1. Data representativeness:

It is important to highlight the tradeoff between data abundance and multidimensionality. Individual-level data provide the richest and most detailed dimensions, but such data are often scarce. For effective analysis and prediction, having abundant data is crucial. Therefore, most of the selected datasets are population-level data. However, some individual patient-level datasets were included because they contained unique and essential dimensions not available in the population datasets. These individual datasets were aggregated to achieve the desired population-level structure. From these individual-level datasets, some metrics of population representativeness can be found. For instance, the vaccine

Table 2
Overview of the data sources.

Aspect	Dataset	Sources	Covered countries	# of instances
Spatiotemporal	Spatiotemporal	Humanitarian data exchange portal (*Welcome - Humanitarian Data Exchange*, 2022), (*Afghanistan: Coronavirus(COVID-19) Subnational Cases - Humanitarian Data Exchange*, 2022)	Afghanistan	2,316,677
		Github https://shorturl.at/LQT78, https://covid19.go.id/	Indonesia	
		Dynamic Covid-19 dashboard for Bangladesh (COVID, DGHS, 2021)	Bangladesh	
		European centre for disease prevention and control (*Data on the weekly subnational 14-day notification rate of new COVID-19 cases*, 2022)	Austria, Italy, Lithuania, Poland, Slovakia, Sweden	
		Naqvi (2021)	Croatia, Denmark, Finland	
		Github (*covid19-data-greece/data/greece/regional at master · Covid-19-Response-Greece/covid19-data-greece*, 2023)	Greece	
		Geohive open data repository (*GeoHive Open Data*, 2023)	Ireland	
		NL Covid-19 geohub repository (*NL COVID-19 Hub*, 2023)	Netherlands	
		Kaggle (*COVID-19 in India | Kaggle*, 2022)	India	
		South Africa provincial breakdown dashboard (*South Africa Provincial Breakdown | Covid-19 South Africa*, 2023)	South Africa	
		Haratian et al. (2021)	United States	
Environment	Weather	World weather online (World Weather Online, 2016)	Countries listed under the Spatiotemporal aspect plus Turkey and Romania	1,337,073
	Air pollution	South Africa air quality information system (SAAQIS) (Gwaze & Mashele, 2018), Central control room for air quality management (CPCBCCR) (*CCR*, 2023)	South Africa, India	157,662
Health	Vaccine hesitancy	Khan, Dabla-Norris, Lima, and Sollaci (2021), YouGov Github repository https://github.com/YouGov-Data/covid-19-tracker	Spain, Netherlands, Japan, Italy, Israel, Germany, France, Denmark, Canada, Australia	430,755
	Comorbidity	Centers for disease control and prevention (CDC) (*Conditions Contributing to COVID-19 Deaths, by State and Age, Provisional 2020–2022 | Data | Centers for Disease Control and Prevention*, 2022)	United States	32,003
Socioeconomic	Ethnicity	Centers for Disease Control and Prevention (CDC) (*Conditions Contributing to COVID-19 Deaths, by State and Age, Provisional 2020–2022 | Data | Centers for Disease Control and Prevention*, 2022)	United States	7,311
	Place of death	Centers for disease control and prevention (CDC) (*Conditions Contributing to COVID-19 Deaths, by State and Age, Provisional 2020–2022 | Data | Centers for Disease Control and Prevention*, 2022)	United States	8,394
	Occupations	Hâncean, Lerner, Perc, Oană, Bunaciu, Stoica, and Ghiță (2022), Sari, Kağan, Karakuş, and Özdemir (2022), Windsor-Shellard and Nasir (2021)	United Kingdom, Romania, Turkey	1918
	Urbanicity	Haratian et al. (2021)	United States	1,572,241

hesitancy dataset and Turkey's occupation-level dataset both consist of individual-level data. As shown in Table 3, the gender ratio in the vaccine hesitancy survey closely aligns with the nationwide gender ratio of the respective countries. In contrast, Turkey's occupation-level dataset shows a male-to-female ratio of 52:100, which significantly differs from the nationwide gender ratio of 100.5:100. While the gender ratios in the vaccine dataset are relatively consistent with national figures, Turkey's occupation dataset deviates considerably. Nevertheless, this dataset was selected because it is one of the few that provides epidemiological data in the context of occupation.

Our ethnicity dataset contains data on ethnic groups in the USA, with Non-Hispanic White, Hawaiian, African-American, and Alaskan

Table 3
Overview of the YouGov vaccine hesitancy dataset (Jones, 2020): distribution of survey participants by country, gender, and population representation.

Country	# of participants	Population (in Million)	Population surveyed (in %)	Male: Female (in population)	Male: Female (in survey)
Spain	54 311	47.26	0.11	96.1:100	95.7:100
Netherlands	19 315	17.34	0.11	98.8:100	94:100
Japan	25 997	124.69	0.02	94.6:100	98.6:100
Italy	54 270	62.39	0.09	95.1:100	91:100
Israel	15 318	8.79	0.17	99.4:100	94:100
Germany	54 513	79.9	0.07	97.4:100	93.2:100
France	54 578	68.08	0.08	93.6:100	86.2:100
Denmark	50 248	5.89	0.85	99:100	95.3:100
Canada	48 372	37.94	0.13	98.8:100	82.3:100
Australia	53 833	25.81	0.21	98.6:100	94.7:100

Native represented in the ratio of 79:0.13:16:1. According to the 2020 USA Census Bureau report, the distribution of these ethnicities in the USA is 57:0.2:12.1:6.1. Although there are slight differences between the two distributions, this dataset was the most comprehensive and disaggregated one available (based on our investigation) that accurately recorded Covid-19 deaths in relation to ethnicity.

As previously mentioned, the tradeoff between data abundance and multidimensionality will inevitably introduce some level of sparsity into the integrated dataset we aim to create. Leveraging paid data sources mitigates this sparsity. Moreover, it is important to note that we are modeling the dataset using a KG, a highly extensible knowledge representation approach; and using RDF as the reference model, our semantic ETL data integration pipeline outlined in Section 5 facilitates the seamless incorporation of new data with minimal effort to extract new data insights. For example, users can integrate misinformation data to analyze how misinformation impacts the spread of a pandemic. The primary reasons for selecting RDF as the reference model are: (1) Its ability to accommodate semi-structured and unstructured data, which can be easily converted into RDF triples using appropriate RDF wrappers. (2) Its schema-relax data integration nature, which also allows integration of data that are not compliant with the predefined schema. In this sense, the integration process aligns with and supports data lake pipelines.

4.2. Modeling the target TBox of CovKG

In this section, we design the MD schema of CovKG to integrate the sources defined in Section 4.1, as depicted in Fig. 4. Since we integrate data from various dimensions and disparate sources, not all data points will align after integration. For instance, the number of deaths of a specific ethnicity in a certain place at a certain time may be available from one data source. The same may be available in case of deaths of patients of a certain comorbidity from another data source. However, to place them in a single cube, we need to know the overlap between the counts i.e., the exact count of deaths in that ethnicity who had that specific comorbidity. This is what we mean, when we say data points will not align. Furthermore, data concerning the same dimensions collected from different sources may be available at different hierarchical levels. To address this issue, we employ data cuboids (Etcheverry et al., 2015). These cuboids are subsets of data cubes and are represented with respect to one or more dimension levels, in contrast to data cubes, which are represented with respect to all dimensions. This approach allows data cuboids to facilitate separation of concerns when analyzing the data.

In Fig. 4, the green cube shapes represent the data cuboids, while the blue rectangles represent the dimensions. An exclusive-or relationship (⊗) between two levels means that the cuboid can include data from either of the levels, but not both simultaneously. Measures of the cuboids are *Total Confirmed Cases* and *Total Deaths*. Among the 13 dimensions, `cdw:GeographyDim` and `cdw:TimeDim` constitute the spatiotemporal perspective; `cdw:TemperatureDim`, `cdw:HumidityDim`, `cdw:WindDim`, `cdw:PrecipitationDim`, and, `cdw:AirPollutionDim` constitute the environmental aspect; `cdw:VaccineHesitancyDim` and `cdw:ComorbidityDim` constitute the health perspective; `cdw:EthnicityDim`, `cdw:PlaceofDeathDim`, `cdw:OccupationDim`, and `cdw:UrbanicityDim` constitute the socioeconomic perspective.

The `cdw:TimeDim` dimension has the `cdw:Calendar` hierarchy, which consists of `cdw:Day`, `cdw:Month`, and `cdw:Year` levels, from the finest to coarsest level. This hierarchy allows the user to see the temporal evolution of the confirmed cases as well as deaths. The `cdw:GeographyDim` dimension has the `cdw:Geography` hierarchy containing `cdw:Admin2` as the finest level, which represents the second administrative level as per Geonames (GeoNames, 2004). The other higher levels are `cdw:Admin1`, `cdw:Country`, and `cdw:Continent`.

The `cdw:TemperatureDim` contains the hierarchy `cdw:Temperature`. It houses two levels (`cdw:ThermalSubtype`, `cdw:Thermaltype`) based on the classification model of Piotrowicz, Ciaranek, Wypych, Razsi, and Mika (2013). Each of the `cdw:HumidityDim`, `cdw:WindDim`, and `cdw:PrecipitationDim` dimensions contains only one level. The `cdw:AirPollutionDim` dimension represents the daily air pollution through various pollutants. This dimension has the `cdw:AirPollution` hierarchy, which contains the `cdw:PollutionLevels` and `cdw:Pollutants` levels. The `cdw:PollutionLevels` level represents the various levels of pollution determined according to the levels depicted in Gadekar (2022). Since Covid-19 is a respiratory disease, it is intuitive that air pollution may have relationships with its epidemiological behavior (Brandt, Beck, & Mersha, 2020; Liang et al., 2020; Travaglio et al., 2021; Wu, Nethery, Sabath, Braun, & Dominici, 2020).

The dimension `cdw:VaccineHesitancyDim` represents the hesitancy to accept vaccines. Its `cdw:Hesitancies` hierarchy has two levels: `cdw:HesitancyScore` and `cdw:VaccineAvailabilityYear`. It is based on the questionnaire used in the research done in Khan et al. (2021). The vaccination intent was determined based on the survey respondents' responses to the question of whether they will get vaccinated if a Covid-19 vaccine becomes available to them in 2021. The respondent can respond with scores 1 (Strongly Agree), 2 (Agree), 3 (Neutral), 4 (Disagree), and 5 (Strong Disagree). The level `cdw:VaccineAvailabilityYear` groups the responses based on year. We have selected vaccine hesitancy as a dimension as it represents people's tendency not to take vaccine, hence assist in the further propagation of the pandemic. In 2019, the WHO listed vaccine hesitancy as one of the top ten threats to global health (*Ten threats to global health in 2019*, 2022).

The dimension `cdw:ComorbidityDim` represents affliction with other diseases alongside Covid-19. Under its `cdw:Diseases` hierarchy, there are two levels- `cdw:Disease` and `cdw:DiseaseType`. Diseases are categorized into three disease types: respiratory diseases, circulatory diseases, and other diseases. Various research has shown that the presence of comorbidities such as diabetes, respiratory diseases, cardiovascular diseases etc., can influence Covid-19's impact on

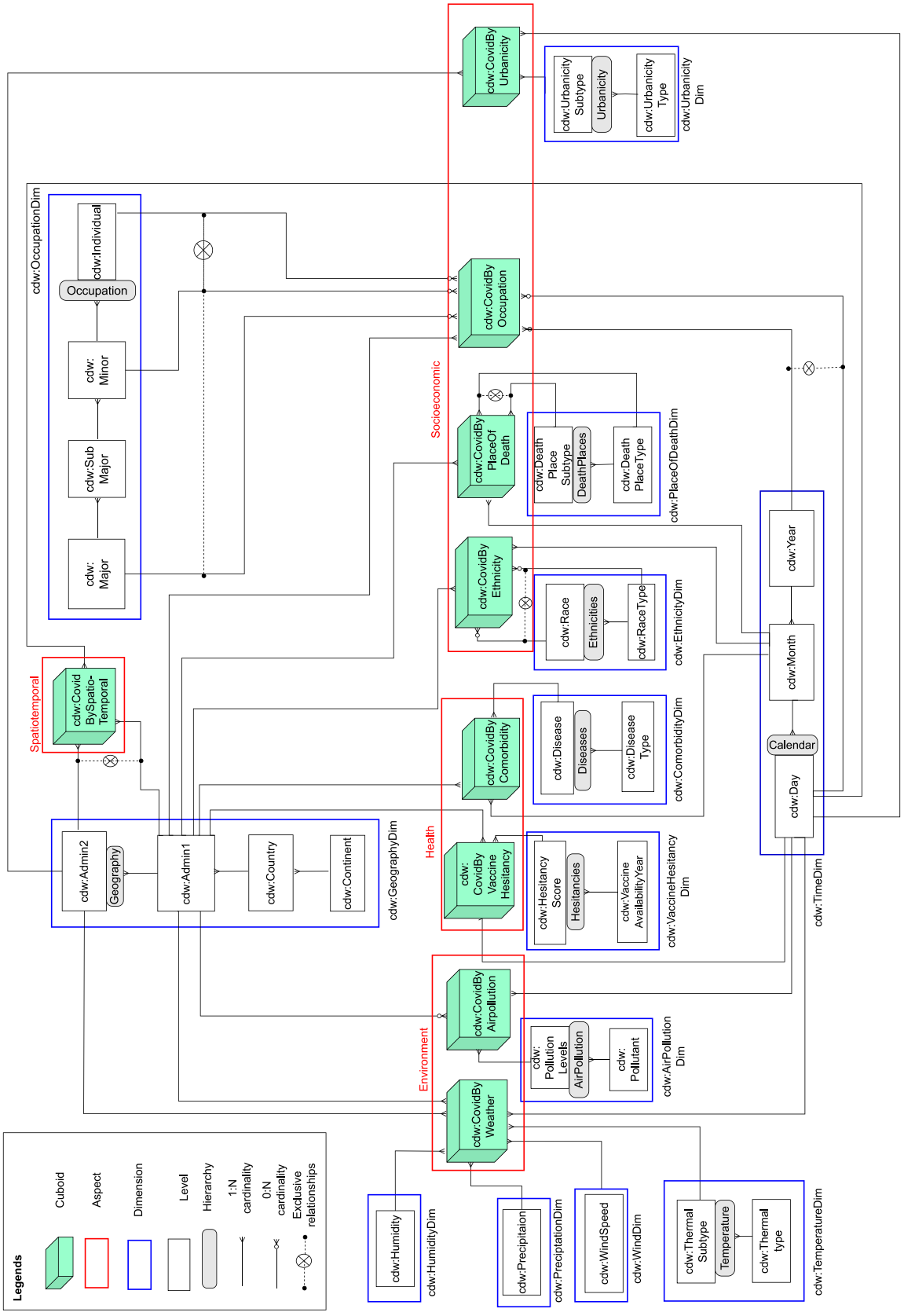


Fig. 4. Multidimensional schema of CovKG.

the patient's body (Guan et al., 2020; Sanyaolu et al., 2020; Wang, Li, Lu, & Huang, 2020). The cdw:EthnicityDim dimension represents the races of the cases in the hierarchy cdw:Ethnicities. This hierarchy holds two levels- cdw:Race and cdw:RaceType. The level cdw:RaceType consists of two instances- Hispanic and Non-Hispanic. Although further classification for Hispanic is not available, Non-Hispanic is divided into five groups, namely, white, black, American Indian/Alaskan native, Asian, and native Hawaiian/Pacific Islander. Multiple researches have shown the existence of ethnic disparity in Covid-19 cases (Ali et al., 2021; Chin-Hong, Alexander, Haynes, & Albert, 2020; Townsend, Kyle, & Stanford, 2020).

Place of death is an important indicator of the availability of healthcare facilities in a geographical area of interest. The dimension cdw:PlaceOfDeathDim indicates the place where the Covid-19 patients died. Under the cdw:PlaceofDeath hierarchy, it has two levels: cdw:DeathPlaceSubtype and cdw:Death . cdw:Death has four types: Healthcare, Homelike, Other, and Unknown. The finer level cdw:DeathPlaceSubtype has three instances both for Healthcare and Homelike, while Other and *Unknown* are not further divided. The three instances in cdw:DeathPlaceSubtype for Healthcare are: Inpatient, Outpatient, and Dead on arrival. The three instances of Homelike are: Decedent's home, Nursing home, and Hospice facility.

Occupations are represented by the cdw:OccupationDim dimension, whose cdw:Occupation hierarchy has four levels as per ISCO-08 (International Labour Office, 2012). Among the four levels, cdw:Individual represents individual occupations and cdw:Minor represents the minor categories which group the individual professions. Similarly, sub-major occupations categorize minor ones, which in turn are aggregated into major occupations. Occupation is an important socioeconomic factor to take into consideration when analyzing Covid-19's epidemiology because occupational distribution can describe people's interaction behavior and hence provide a precursor to the transmission path of the disease.

The dimension cdw:UrbanicityDim represents the urban–rural classification of a geographical area as per the urban–rural classification scheme for counties (Ingram & Franco, 2014). The cdw:UrbanicitySubtype level is composed of six instances: Large central metro, large fringe metro, medium metro, small metro, micropolitan, and noncore. Among them, the first four belong to the metropolitan (urban) category and the last two belong to the noncore (rural) category. Noncores and metropolitans constitute the level cdw:UrbanicityType.

5. Generation of CovKG

In this section, we will elaborate on how the task of generating CovKG is accomplished using an ETL pipeline.

In the *Extraction phase*, we first extract data from different sources as described in Section 4.1, then cleanse and format those data to conform with the target TBox. This cleansing and formatting tasks include extraction of micro data, elimination of aggregated data, conversion from other formats to CSV, adding unique ids, filtering out irrelevant and noisy data, and so on. Then, the *Transformation* phase semantically transforms the extracted data according to the semantics encoded in the target TBox and generates CovKG, which is in turn loaded into the triple store Openlink Virtuoso (Erling, 2012) in the *Load phase*.

The *Transformation* process unfolds in four steps: (1) The *Target TBox generation* step implements the target TBox defined in Section 4.2. (2) The *Source TBox generation* step generates TBoxes from the data sources. (3) The *SourceToTarget mappings generation* step establishes mappings between the source and target TBoxes. (4) Then, the *Target ABox generation* process generates assertions consistent with the target TBox. Below, we provide a detailed description of each of these steps.

5.1. Target TBox generation

The purpose of this step is to represent the target TBox as outlined in Section 4.2 using the constructs provided by RDFS, OWL, and QB4OLAP (as defined in Section 2) alongside the RDF model. Users have the flexibility to create a TBox with MD semantics either manually or by utilizing tools such as Protege (Musen, 2015), WebVOWL (Lohmann, Link, Marbach, & Negru, 2015), or PoolParty (Knap et al., 2018). Listing 1 demonstrates a part of the target TBox annotated with QB4OLAP constructs. In this listing, cdw: represents the namespace of the data cube, which is https://bike-csecu.com/datasets/CovKG/cdw.ttl#. The cubic structure of the dataset cdw:SpatioTemporalDataset is defined by cdw:SpatioTemporalCuboid (lines 5–12), which contains the cdw:Admin2 and cdw:Admin1 levels of the cdw:GeographyDim dimension, and cdw:Day of the cdw:TimeDim dimension. Both cdw:Admin2 and cdw:Admin1 are included in this cuboid, as some countries have data available only at the first administrative level, while others have data at the second administrative level.

The cdw:TimeDim dimension contains the cdw:calendarHierarchy hierarchy which is composed of cdw:Day, cdw:Month, and cdw:Year levels (lines 14–21). A hierarchy step of cdw:calendarHierarchy (lines 23–28) represents that cdw:Day is related to its parent level cdw:Month through the rollup property defined by cdw:inMonth (line 28). The level cdw:Day is defined as an instance of the qb4o:LevelProperty class, and it contains a set of attributes (lines 30–37). cdw:dayID is a level attribute (lines 35–37) and cdw:inMonth is defined as a roll-up relation (lines 39–41). The measures cdw:Confirmed and cdw:Deaths are defined using the qb:MeasureProperty class (lines 43–48). They are defined as decimals to facilitate aggregation functions like average, which may return floating point values.

5.2. Source TBox generation

After creating the target TBox of CovKG, the next task is to populate CovKG from the available data sources. In order to accomplish this, we must establish mappings between the target and source constructs at the TBox level. Therefore, it is essential to derive TBoxes from the existing sources and augment them with OWL and RDFS constructs. Various vocabularies/tools, such as R2RML mapping (World Wide Web Consortium et al., 2012), Direct mapping (Arenas et al., 2012), and NonSemanticToTBoxDeriver (Nath et al., 2017), can be used for this purpose. In this study, we utilize NonSemanticToTBoxDeriver. This tool is employed to extract conceptual information embedded in non-semantic structured data and transform it into a semantic form. Listing 2 displays the generated source TBox from the spatiotemporal dataset. Here, the table name is used as an OWL class and the attributes are considered as OWL datatype properties. The onto: namespace is used to create semantic counterparts of the various elements of the source data.

5.3. Source-to-target mappings generation

Source data can be heterogeneous in nature. To handle this situation, sources should be mapped to the target at the TBox level. The communication between the source and target is materialized in the form of intermediate mapping definitions that assist complex data flows between the sources and target. As source TBoxes are generated for both dimension tables and fact tables, the mappings are to be generated for both as well. Source-to-target mappings of dimension tables are used later to produce the semantic assertions of the level instances. Similarly, source-to-target mappings of fact tables are used later to create semantic assertions of the fact tables. Listing 3 shows mapping definitions between different constructs of onto:SpationtemporalFact and cdw:spatioTemporalDataset. The mapping file are annotates with Source-to-Target Mapping (S2TMAP) vocabulary (Deb Nath,

```
#DATASETS
cdw:SpatioTemporalDataset a qb:DataSet;
  qb:structure cdw:SpatioTemporalCuboid.
#CUBOIDS
cdw:SpatioTemporalCuboid a qb:DataStructureDefinition;
  dct:conformsTo <http://purl.org/qb4olap/cubes>;
  qb4o:isCuboidOf cdw:COVID_DW;
  qb:component [ qb:measure cdw:Confirmed; qb4o:aggregateFunction qb4o:sum];
  qb:component [ qb:measure cdw:Deaths; qb4o:aggregateFunction qb4o:sum];
  qb:component [ qb4o:level cdw:Admin1; qb4o:cardinality qb4o:OneToMany];
  qb:component [ qb4o:level cdw:Admin2; qb4o:cardinality qb4o:OneToMany];
  qb:component [ qb4o:level cdw:Day; qb4o:cardinality qb4o:OneToMany].
#DIMENSIONS
cdw:TimeDim a qb:DimensionProperty;
  rdfs:label "Time Dimension"@en;
  qb4o:hasHierarchy cdw:CalendarHierarchy.
#HIERARCHIES
cdw:CalendarHierarchy a qb4o:Hierarchy;
  rdfs:label "Calendar Hierarchy"@en;
  qb4o:inDimension cdw:TimeDim;
  qb4o:hasLevel cdw:Day, cdw:Month, cdw:Year.
#HIERARCHY STEPS
_:hs16 a qb4o:HierarchyStep;
  qb4o:inHierarchy cdw:CalendarHierarchy;
  qb4o:childLevel cdw:Day;
  qb4o:parentLevel cdw:Month;
  qb4o:pcCardinality qb4o:OneToMany;
  qb4o:rollup cdw:inMonth.
#LEVELS
cdw:Day a qb4o:LevelProperty;
  rdfs:label "Day"@en;
  qb4o:hasAttribute cdw:dayID, cdw:dayName, cdw:inMonth;
  rdfs:range cdw:Day.
#ATTRIBUTES
cdw:dayID a qb4o:LevelAttribute;
  rdfs:label "Day ID"@en;
  rdfs:range xsd:string.
#ROLLUP RELATIONSHIPS
cdw:inMonth a qb4o:LevelAttribute, qb4o:RollupProperty;
    rdfs:label "Rollup property to roll up from day to month"@en;
    rdfs:range onto:Month.
#MEASURES
cdw:Confirmed a qb:MeasureProperty;
  rdfs:label "Total Confirmed Cases"@en;
  rdfs:range xsd:decimal.
cdw:Deaths a qb:MeasureProperty;
  rdfs:label "Total Deaths"@en;
  rdfs:range xsd:decimal.
```

Listing 1: Target TBox defining spatiotemporal cuboid and the time dimension. Prefixes are omitted due to space constraints.

```
onto:SpatiotemporalFact  a   owl:Class.
onto:adm2ID  a   owl:DatatypeProperty;
        rdfs:domain  onto:SpatiotemporalFact;
        rdfs:range  xsd:string.
onto:Death  a  owl:DatatypeProperty;
        rdfs:domain  onto:SpatiotemporalFact;
        rdfs:range  xsd:decimal.
onto:adm1ID  a   owl:DatatypeProperty;
        rdfs:domain  onto:SpatiotemporalFact;
        rdfs:range  xsd:string .
onto:Confirmed  a   owl:DatatypeProperty;
        rdfs:domain  onto:SpatiotemporalFact;
        rdfs:range  xsd:decimal.
onto:dayID  a   owl:DatatypeProperty;
        rdfs:domain  onto:SpatiotemporalFact;
        rdfs:range  xsd:string.
```

Listing 2: Source TBox for the fact table SpatiotemporalFact.

```
#Dataset mapping
cdw:spatiotemporalfacts_COVID_Schema a map:Dataset ;
        map:source  '/SpatiotemporalFacts';# source location
        map:target  '/COVID_Schema'.# target location
#Concept mapping
cdw:SptempFact_SpTempDataset  a map:ConceptMapper;
        map:sourceConcept     onto:SpatiotemporalFact;
        map:targetConcept     cdw:SpatioTemporalDataset.
        map:dataset   cdw:spatiotemporalfacts_COVID_Schema;
        map:iriValue  "CONCAT(onto:adm1ID,CONCAT(_,CONCAT(onto:adm2ID,CONCAT(_,onto:dayID))))";
        map:iriValueType     map:Expression;
        map:matchedInstances "All";
        map:relation          skos:exact;
#Property mapping
map:PropMap_Confirmed_Confirmed a map:PropertyMapper ;
        map:ConceptMapper        cdw:SptempFact_SpTempDataset;
        map:sourceProperty       onto:Confirmed;
        map:sourcePropertyType   map:SourceProperty;
        map:targetProperty       cdw:Confirmed.
map:PropertyMapper_01_dayID_day a  map:PropertyMapper;
        map:ConceptMapper        cdw:SptempFact_SpTempDataset;
        map:sourceProperty       onto:dayID;
        map:sourcePropertyType   map:SourceProperty;
        map:targetProperty       cdw:Day.
```

Listing 3: Source-to-Target mapping file of `onto:SpationtemporalFact` and `cdw:spatioTemporalDataset`.

Romero, Pedersen, & Hose, 2022): an OWL-based mapping vocabulary.

In S2TMAP, a property-level mapping is nested within a concept-level mapping, which is further encapsulated within a mapping dataset. A mapping dataset is defined as an instance of `map:Dataset`, which captures the references of the source and target TBoxes (lines 2–4). A concept-mapping outlines the correspondence between a source and a target concept (lines 6–13). The source and target concepts are defined using the `map:sourceConcept` and `map:targetConcept` properties. The linkage between a concept-mapping and its mapping dataset is established through the `map:dataset` property. The properties `map:iriValue` and `map:iriValueType` signify that values of `onto:admID, onto:adm2ID`, and `onto:dayID` are concatenated to generate unique IRIs for the observations of `cdw:SpationTempo ralDataset`. The "All" value of `map:matchedInstances` indicates that all source instances are mapped.

A source and target property are mapped using the `map:Property Mapper` (lines 15–19). The connection between a property-mapping and its corresponding concept-mapping is established via `map:concep tMapper`. The specification of the target property within the property-mapping is done using `map:targetProperty`, and this target property can be associated with either a source property or an expression. In this specific instance, the target property `cdw:Confirmed` is mapped to the source property `onto:Confirmed`.

5.4. Target ABox generation

Using the target TBox, along with the source datasets (extracted and cleansed ones) and source-to-target mapping definitions as inputs, this *Target ABox generation* process generates the target ABoxes from the source datasets based on the semantics specified in the target TBox. In QB4OLAP, dimensional data is physically stored in levels, where each level member is identified by a unique IRI and is semantically linked with its relevant level attributes and roll-up properties.

In Listing 4, a level member of the `cdw:Day` level, with an IRI value of `day:1`, is implemented using the `qb4o:LevelMember` class. This level member has the attributes `cdw:dayID` with a value of "1", `cdw:dayName` with a value of "2020-01-01", and `cdw:inMonth` with a value of `month:1`. The `cdw:inMonth` represents the relationship of the `cdw:Day` level with its parent level `cdw:Month`. Using the property `qb4o:memberOf`, the relationship of the level member to its containing level is defined. Moreover, the `owl:sameAs` property is used to link to an external data resource. In this case, the resource is the Wikidata entity `wd:Q57396575`, which is the Wikidata entry for January 1, 2020. QB4OLAP employs an observation (an instance of `qb:Observation`) to depict a fact (line 1 in Listing 5). An observation is identified by a distinct IRI and is semantically enriched by combining multiple members from different levels, integrating values for various measure properties. Listing 5 depicts that the dataset of the observation is `cdw:SpationTemporalDataset` and its cuboid structure consists of `cdw:Admin1, cdw:Admin2` and `cdw:Day` levels and two measures: `cdw:Confirmed` and `cdw:Deaths`.

6. Description of CovKG

In this section, we describe CovKG from the perspective of both dimensions and facts. Additionally, we provide an overview of the embedded links in CovKG leading to external datasets.

6.1. Dimension and fact overview

Table 4 provides an overview of the dimensions of CovKG. It shows the aspect the dimension represents, the total number of level instances, the number of level attributes. In total, CovKG has 28 levels, 87 level attributes, and 6210 level members. Nine separate Turtle ABox files are created for nine cuboids. These files are then concatenated with the respective level ABoxes of the dimensions used in the fact tables. Table 5 sheds light on the size measurements of CovKG. The raw source ABoxes are available in CSV format, containing a lot of noise data such as irrelevant information not useful for the purpose of this study, negative values, and blank values. The raw source ABoxes are processed and cleaned, after which their sizes reduce significantly, as can be seen in the table.

Another reason for this reduction in size was the utilization of dimension tables. Dimension information in the fact tables, such as names and labels, or floating-point sensor data, are replaced by integer indices pointing to the relevant dimension tables. This information is instead placed in the dimension table, which the fact table can refer to when needed. Additionally, the arrangement of dimension tables in hierarchical levels allows for the performance of OLAP operations on these fact tables. Furthermore, floating-point data of air pollution

```
day:1  a   qb4o:LevelMember;
 cdw:dayID  "1";
 cdw:dayName  "2020-01-01";
 cdw:inMonth  month:1;
 qb4o:memberOf  cdw:Day;
 owl:sameAs  wd:Q57396575.
```

Listing 4: A level member of the `cdw:Day` level in the target ABox.

```
stempd:_1224_52  a   qb:Observation;
   cdw:Confirmed  "8096";
   cdw:Deaths  "270";
   cdw:Admin1   _:b;
   cdw:Admin2   adm2:1224;
   cdw:Day   day:52;
   qb:dataset  cdw:SpatioTemporalDataset.
```

Listing 5: An observation of the `cdw:SpatioTemporalDataset` dataset in the target ABox.

Table 4
A quantitative overview of the dimensions present in CovKG.

Aspect	Dimension	# of instances	# of attributes
Spatiotemporal	cdw:GeographyDim	4,327	11
	cdw:TimeDim	1,135	8
Sub Total		**5,462**	**19**
Environment	cdw:TemperatureDim	15	11
	cdw:HumidityDim	4	4
	cdw:WindDim	4	4
	cdw:PrecipitationDim	3	4
	cdw:AirPollutionDim	42	9
Sub Total		**68**	**32**
Health	cdw:VaccineHesitencyDim	21	5
	cdw:ComorbidityDim	24	5
Sub Total		**45**	**10**
Socioeconomic	cdw:EthnicityDim	7	5
	cdw:PlaceofDeathDim	10	5
	cdw:OccupationDim	610	11
	cdw:UrbanicityDim	8	5
Sub Total		**635**	**26**
Grand Total		**6,210**	**87**

Table 5
Overview of size metrics of the data cuboids.

Semantic cuboid	Source ABox raw size (in MB)	Source ABox processed size (in MB)	# of observations (in million)	# of RDF triples (in million)	Target ABox size (in MB)
Spatiotemporal	838.9	377.4	2.32	16.25	1,719.5
Weather	102.3	26.9	1.34	14.75	1,646.1
Air Pollution	12.8	3.1	0.16	1.14	121.7
Vaccine hesitancy	179.9	0.104	0.002	0.054	5.2
Comorbidity	65.6	2.6	0.032	0.26	27.9
Ethnicity	1.4	0.106	0.0073	0.095	9.6
Place of death	2.7	0.121	0.0084	0.1	10.9
Occupations	4.6	0.035	0.0019	0.06	6.1
Urbanicity	705.4	29.5	1.57	11.04	1,185.0
Total	**1,913.6**	**439.87**	**5.44**	**43.75**	**4,732.1**

and weather dimensions are replaced by categorical data, which also contributes to the reduction in size after processing.

The number of RDF triples in Table 5 is significantly larger than the number of observations. This is due to the fact that for each observation, multiple RDF triples are generated. For instance, if an observation has five attributes, containing three dimension attributes and two measure attributes, then there will be seven RDF triples representing that observation in the fact table.

Concatenating the respective dimension level ABox files to the fact ABox files also increase the number of triples. The target ABox sizes are

Table 6
Number of links to external datasets among level instances and the programmatic time taken to link.

Level	Number of links	Processing time (sec)
cdw:Day	1,096	1.712
cdw:Month	36	0.751
cdw:Year	3	0.726
cdw:Individual	872	0.972
cdw:Minor	260	0.496
cdw:Submajor	86	0.482
cdw:Major	20	0.42
cdw:Continent	14	2.292
cdw:Country	44	2.333
cdw:Admin1	922	2.504
cdw:Admin2	7,664	6.414
Total	**11,017**	**19.102**

relatively large. As can be seen in Table 5, the spatiotemporal, weather, and urbanicity datasets have gigabyte-scale sizes. This is primarily because the spatiotemporal data contains finer spatial data in the form of second-level administrative unit and first-level administrative unit data for twenty one countries. Spatiotemporal data was one of the most readily available types of data collected in this study. On the temporal side, day-level data is available, contributing to the increase in size. Additionally, the inclusion of daily confirmed and death data for 3143 USA counties significantly contributed to the overall data size.

Urbanicity data was also based on the USA's spatiotemporal data. The weather data was available for all countries at the same fine level as the spatiotemporal data and had more dimension fields than the spatiotemporal data. Temperature, precipitation, and humidity data were not available for the USA, which is why the weather dataset is still smaller in size than the spatiotemporal dataset. The core reason behind the exponential size of the RDF data is that Turtle requires more text characters to represent data than tabular data such as CSV. However, this size tradeoff is reasonable, as the dataset achieves the ability to infer new knowledge based on the available information. Moreover, it gains the capability to be linked to larger knowledge networks to mine further insights, utilizing techniques such as federated query.

6.2. Linking CovKG to external datasets

Linked open datasets contain references to similar elements across other external datasets, allowing for the sharing and reusability of previous knowledge. This also helps in avoiding the inclusion of redundant data, contributing to maintaining scalability. The linking can be done to the concepts in the target TBox as well as level instances in the target ABox (Nath, Seddiqui, & Aono, 2014). CovKG is linked to a total of four reputable external KGs. This is achieved using the OWL property `owl:sameAs`, as demonstrated in Listing 4 and Listing 6. In the target TBox, all the levels of `cdw:GeographyDim`, `cdw:TimeDim`, `cdw:AirPollutionDim`, `cdw:PrecipitationDim`, `cdw:WindDim`, `cdw:HumidityDim`, as well as the level `cdw:Race` of `cdw:EthnicityDim` are linked to Wikidata (Vrandečić & Krötzsch, 2014) and DBpedia (Auer et al., 2007). Moreover, levels under `cdw:GeographyDim` are linked to the Geonames KG (GeoNames, 2004). Geonames is renowned for collecting and presenting geographical information at highly fine levels in semantic form. Demonstrations of some TBox links are shown in Listing 6 under the comment 'Level in the target TBox' (lines 2–7).

In the ABoxes, level instances of the `cdw:GeographyDim` dimension are linked to Wikidata and Geonames. Level instances of the `cdw:TimeDim` dimension's levels are linked to Wikidata. The level instances of the `cdw:OccupationDim` dimension are linked to Wikidata and the European Skills, Competences, qualifications and Occupations (ESCO) ontology (De Smedt, le Vrang, & Papantoniou, 2015). The ESCO ontology makes the ISCO-08 occupation taxonomy available in semantic form. Demonstrations of some ABox links are shown in Listing 6 under the comment 'Level instances in the target ABox' (lines 9–17).

This linking process is implemented using the RDFlib Python library (Krech, 2006). Initially, the IRIs of the level members to external KGs, such as Wikidata (Vrandečić & Krötzsch, 2014), Geonames (GeoNames, 2004), and ESCO (De Smedt et al., 2015), are collected through SPARQL queries on the Wikidata SPARQL endpoint. The query results are imported as CSV files and undergo pre-processing to eliminate duplicates and irrelevant data. Finally, RDFlib is employed to link these IRIs and the ABox triples of the corresponding dimension levels or members. The number of links to the level instances as well as their respective programmatic linking times are reported in Table 6. In summary, CovKG is linked to 10,951 external resources and the total linking time is 19 s.

6.3. Availability

The dump files of CovKG can be found at http://bike-csecu.com/datasets/CovKG, and CovKG was stored in the OpenLink Virtuoso Triplestore. Users can remotely access CovKG through the SPARQL endpoint at https://covkg.bike-csecu.com/sparql/ and write their own SPARQL queries based on their requirements to obtain answers. To verify the correctness and conduct comparative analysis of CovKG, we developed a set of competency questions (refer to Table 8 and Table 10). All these competency questions are translated into equivalent SPARQL queries to retrieve the answers from CovKG. The set of competency and correctness queries can be accessed, posed to the repository, and answered through a user interface available at https://bike-csecu.com/datasets/CovKG/query. We also provide an interactive OLAP interface, available at https://bike-csecu.com/datasets/CovKG/analytical-interface, allowing users to create their OLAP queries using GUI components and retrieve the answer by posing the query to the related graphs. The OLAP interface is described in Section 7.2.1.

7. Experimental evaluation

In this section, we discuss experiments conducted on CovKG to evaluate its performance. First the ETL performance is measured by the ETL runtime. After that, we make the qualitative assessment of CovKG. Finally, we present some interesting analytical findings.

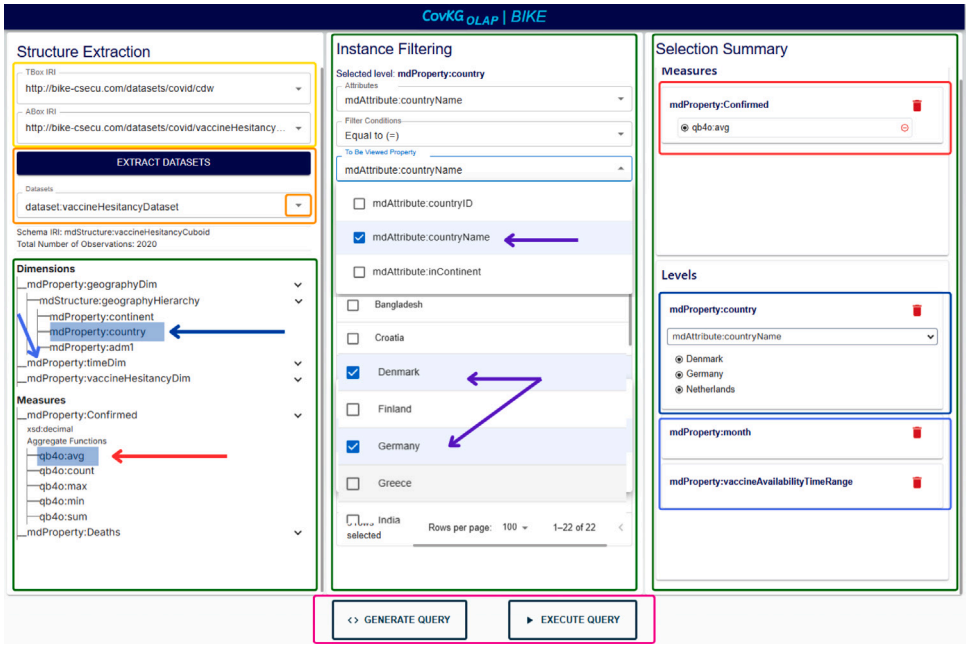
7.1. ETL performance overview

This section discusses the ETL time performance. The ETL process was executed on a computer equipped with an Intel(R) Core(TM) i5-8400 CPU, operating at a speed of 2.81 GHz with 8 GB of RAM. The system runs on Windows 10 Pro (64-bit).

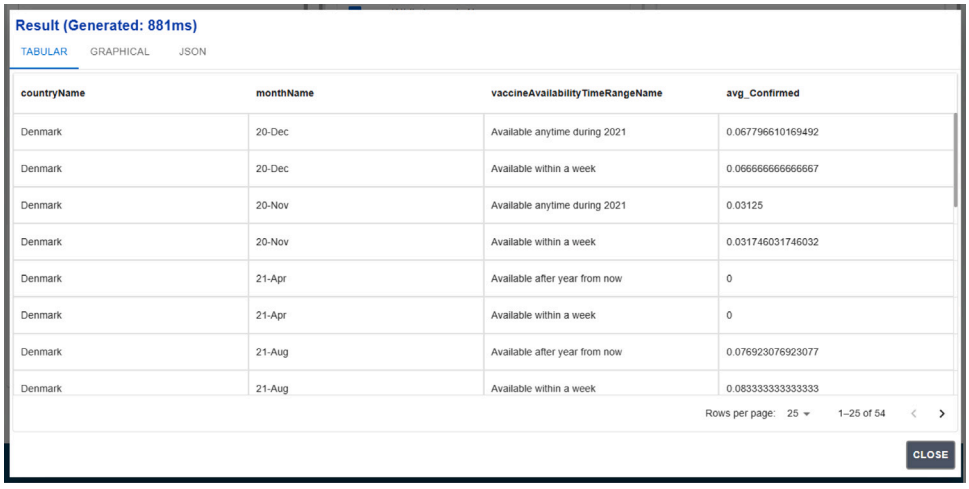
Table 7 shows the ETL time performance (measured in seconds) in outputting CovKG. We do not record the time for the extraction phase as it depends on users' expertise, internet speed, API performance.

```
#Level in the target TBox
cdw:Admin1 a qb4o:LevelProperty;
        owl:sameAs wiki:Q10864048, geonames:A.ADM1, dbpedia:First- level_administrative_division.
cdw:Day  a qb4o:LevelProperty;
         owl:sameAs wiki:Q573, dbpedia:Day.
cdw:Humidity a qb4o:LevelProperty;
               owl:sameAs wiki:Q180600, dbpedia:Air_humidity.
#Level instances in the target ABox
adm1:1 a qb:LevelMember;
       cdw:adm1Name    "Badakhshan";
       owl:sameAs wiki:Q165376, geoname:1147745.
indiOcc:0110 a qb:LevelMember;
       cdw:individualOccupationName    "Commissioned Armed Forces Officers";
       owl:sameAs wiki:Q108305412, esco:C0110.
day:1 a qb:LevelMember;
       cdw:dayName   "2020-01-01";
       owl:sameAs wiki:Q57396575.
```

Listing 6: Examples of links to external datasets.



(a) Loading the Vaccine Hesitancy cuboid in.



(b) Results of a slice query.

Fig. 5. Enabling business intelligence on CovKG.

Table 7
ETL program time taken (seconds) by the ETL process for each cuboid.

Semantic cuboid	TBox generation	Source to target mapping	ABox generation	RDF loading	Total (per cuboid)
Spatiotemporal	6.63	1.29	123	227.67	358.59
Weather	10.65	1.23	150	960	1,121.88
Air Pollution	7.68	1.63	8	39.97	57.28
Vaccine Hesitancy	7.48	1.71	1.3	5.62	16.11
Comorbidity	7.71	1.19	5	10.58	24.48
Ethnicity	7.56	1.28	1	5.55	15.39
Place of death	7.53	1.34	2	5.22	16.09
Occupations	8.54	1.30	1.73	1.801	13.371
Urbanicity	7.5	1.22	77	838.36	924.08
Total (per phase)	**71.28**	**12.19**	**369.03**	**2,094.771**	Grand Total =2,547.271

Table 8
The competency queries designed for qualitative analysis of CovKG.

Query no.	Competency query statement
Q1	What is the name and population of the level x (example:district) with max confirmed cases on date y(example January 1, 2021) in location z (example:Bangladesh)?
Q2	What is the number of deaths among occupation x (example: Engineers) in location y (example: Romania) in the year z (example : 2020) at hot temperature?
Q3	What disease comorbidity has the highest number of deaths during month x (example:February) of year y(example:2020)?
Q4	Which has more infections of Covid-19? Urban(metropolitan) or Rural((non-metropolitan)?
Q5	Which countries have the strongest vaccine hesitancy to vaccines available after one year?
Q6	How many deaths occurred in homelike environments in month x (example:January) of year y (example:2021)?
Q7	What kind of thermal weather has most number of confirmed and/or death on date x (example:January 1, 2021) in location y (example: Feni,Chittagong, Bangladesh)?
Q8	What kind of humidity has most number of confirmed and/or death on date x (example:January 1, 2021) in location y (example: Feni,Chittagong, Bangladesh)?
Q9	What kind of precipitation has most number of confirmed and/or death on date x (example:January 1, 2021) in location y (example: Feni,Chittagong, Bangladesh)?
Q10	What kind of windspeed has most number of confirmed and/or death on date x (example:January 1, 2021) in location y (example: Feni,Chittagong, Bangladesh)?
Q11	How many patients died of Covid-19 in Asia in region of hazardous air pollution in 2020?
Q12	What race has the highest number of deaths in 2021?
Q13	What is the total number of confirmed cases in medical professions?

Rather, the steps which can be calculated fairly are recorded. It can be seen that larger fact ABoxes such as the spatiotemporal, urbanicity, and weather take longer to pass through the ETL process. The greatest proportion is taken up at RDF loading time, where the facts are loaded to the triple store. We use Openlink Virtuoso (Erling, 2012) as triple store because of generating fast query results, simple interface, and code correction ability.

The source TBox generation time is shown here. The target TBox generation's time was not provided here because it was a manual process where the authors had to carefully design the structure. The entire ETL process takes 2547.271 s, around 42 min. The majority of the time is spent on RDF loading, primarily due to loading RDF graphs into the triple store. Among the cuboids, the weather cuboid takes the longest time, as it contains the most attributes.

7.2. Qualitative analysis

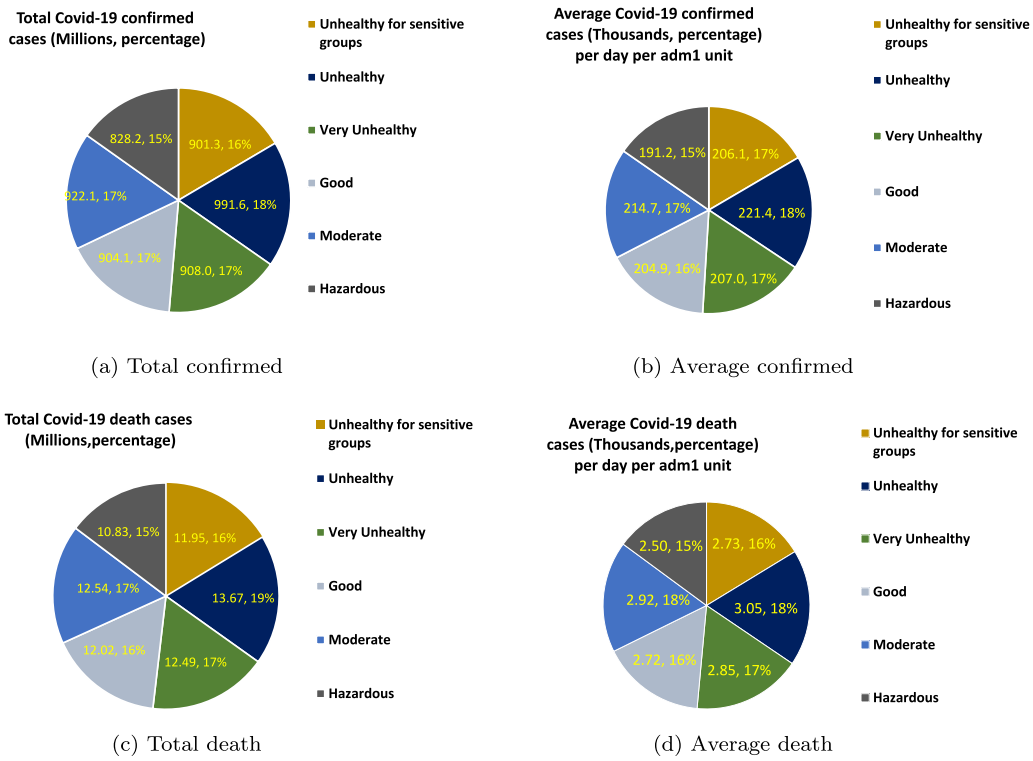
The previous subsection focused on the quantitative performance evaluation of CovKG. In this section, we assess the quality of CovKG in terms of its business analytical capabilities, performance with other repositories, and correctness.

7.2.1. Enabling business analytics

After generating CovKG, we evaluate its business analytical capabilities. This assessment focuses on whether CovKG has become OLAP-compatible, and to do so, we provide $CovKG_{OLAP}$ - an online available application to run OLAP operations over CovKG annotated with MD semantics. Therefore, if CovKG can be loaded into the $CovKG_{OLAP}$ system, OLAP operations can be conducted, and results can be generated, then it confirms that CovKG is ready for business analytics.

In Fig. 5, we demonstrate how CovKG is enabled for business analytics using the $CovKG_{OLAP}$. For brevity, we illustrate the business analytics enabling only one of the nine fact ABoxes, which is the Vaccine Hesitancy ABox. Fig. 5(a) shows that the Vaccine Hesitancy cuboid is successfully loaded into the tool. To load the CovKG, a user needs to select the TBox and the ABox from the dropdown menus (marked by a yellow rectangle). After selecting the files and clicking the *EXTRACT DATASETS* button, the system loads them if the ABox file is OLAP compatible. The upper left part of Fig. 5(a) shows that the target TBox and ABox files have loaded successfully (marked by the yellow rectangle). After loading them, the user selects the Vaccine Hesitancy cuboid from the drop-down list using the arrow icon (marked by the orange square) to extract the cuboid's structure. The *Visualization* panel on the left displays dimensions, hierarchies, levels, and measures in a tree view. Users can expand the tree view by clicking on the titles, as shown by the blue arrow. Levels, measures, and aggregation functions are selected from this panel.

The *Instance Filtering* panel in the middle shows the available attributes when a level is selected. For instance, when the `cdw:Country` level (marked by the purple arrow) is selected, attributes like `cdw:countryName` are displayed (as shown by the purple arrow) for the user to choose. Additionally, one can use checkboxes to select instances for



(a) Total confirmed (b) Average confirmed

(c) Total death (d) Average death

Fig. 6. Covid-19 situation with respect to nitrogen dioxide pollution in India and South Africa shown in millions and percentage for total and thousands and percentage for average per day per Admin1 unit.

Table 9
Assessing the comparative functionality of CovKG against prominent data sources by asking each the thirteen competency questions, to which they answer in yes, partially or no.

Competency query no.	CovKG	Bangladesh dashboard (COVID, DGHS, 2021)	CDC dashboard (Centers for Disease Control and Prevention, 2023)	WHO dashboard (*WHO Coronavirus (COVID-19) Dashboard*, 2023)	World Bank dashboard (World Bank, 2023)	Worldometer (Worldometer, 2020)
Q1	Yes	Partially	Partially	No	No	No
Q2	Yes	No	No	No	No	No
Q3	Yes	No	Yes	No	No	No
Q4	Yes	No	Yes	No	Yes	No
Q5	Yes	No	Partially	No	Partially	No
Q6	Yes	No	Yes	No	No	No
Q7	Yes	No	No	No	No	No
Q8	Yes	No	No	No	No	No
Q9	Yes	No	No	No	No	No
Q10	Yes	No	No	No	No	No
Q11	Yes	No	No	No	No	No
Q12	Yes	No	Yes	No	No	No
Q13	Yes	No	No	No	No	No

slice and dice operations. The rightmost *Selection Summary* panel shows the selected level, their associated attributes, level instances, measures, and aggregate functions. For the example in the figure, the `qb4o:avg` function is selected for the measure `cdw:Confirmed` (marked by the red arrow). Fig. 5(b) demonstrates the result of a slice OLAP operation with averaging as the aggregate function, displaying data only for Denmark, Germany, and the Netherlands along the geography dimension.

$CovKG_{OLAP}$ is accessible at https://bike-csecu.com/datasets/CovKG/analytical-interface/, allowing users to perform various OLAP operations on CovKG and obtain results in tabular, graphical, or JSON formats.

7.2.2. Comparative analysis of functionality in relation to leading data repositories

We compare the functionality of CovKG with that of other data repositories by first formulating a set of competency queries. These competency queries are designed to assess the ability of the datasets to answer questions with multiple aspects (Nath, Das, Das, & Raihan, 2024). The questions are listed in Table 8. This evaluation is conducted in comparison to the responses of well-known repositories: Worldometer (Worldometer, 2020), WHO Dashboard (*WHO Coronavirus (COVID-19) Dashboard*, 2023), World Bank Dashboard (World Bank, 2023), Dynamic Dashboard for Bangladesh (COVID, DGHS, 2021), and CDC COVID Data Tracker Dashboard (Centers for Disease Control and Prevention, 2023). The responses are summarized in Table 9.

The comparison reveals that CovKG can answer all the competency questions. In contrast, the Bangladesh Dashboard, World Bank

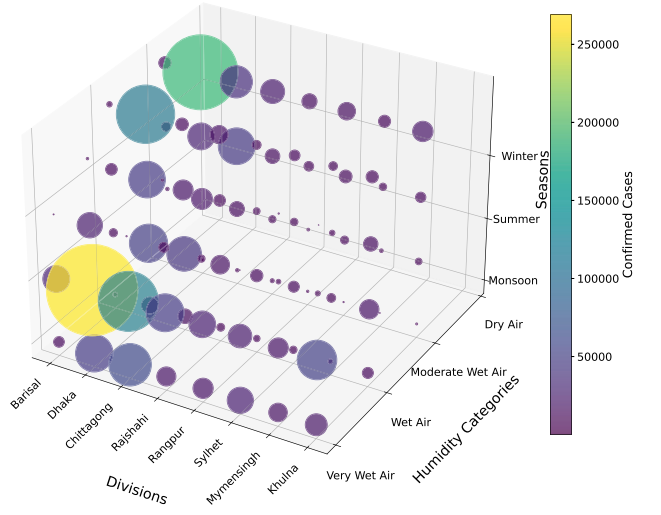


Fig. 7. Confirmed cases in Bangladesh divisions by seasons and humidity levels during 2020 and 2021.

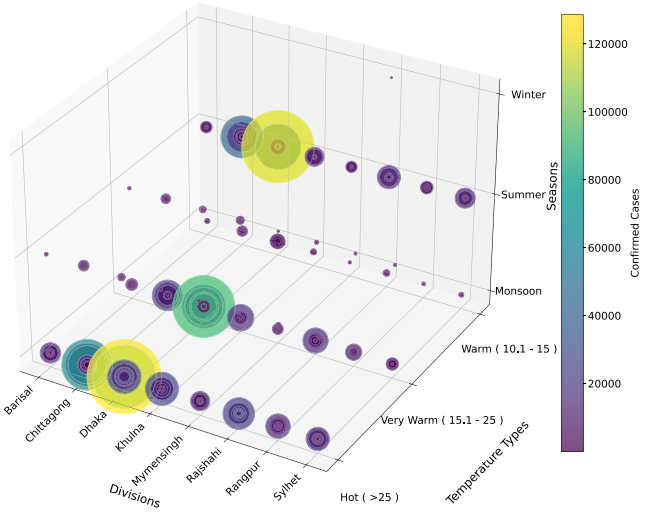


Fig. 8. Confirmed cases in Bangladesh divisions by seasons and temperature variations during 2020 and 2021.

Dashboard, and CDC Dashboard can only answer 3.84%, 11.54%, and 39.23%, respectively. The WHO and Worldometer Dashboards cannot answer any of these questions. When we indicate "partially", it means that the dashboard can address some aspects of the question but not the entire inquiry. For instance, the Bangladesh and CDC Dashboards can only provide district and division-level daily data but lack population information.

Interestingly, CovKG does not contain population information in any of its fact or dimension tables. Yet, it can fully answer Q1 with the assistance of a federated query. This is made possible because CovKG is linked to external KGs through the `owl:sameAs` property, as demonstrated in Listing 4. Q5 can be answered by the CDC and World Bank Dashboards, as they provide extensive vaccine hesitancy-related information. However, they do not address the question regarding the availability timeframe of 'after one year'. This query is valuable in assessing people's psychological attitudes toward vaccination. It indicates whether people intend to take vaccine given it has been tested properly, or whether they do not prefer it under any circumstances.

7.2.3. Correctness

In our study, data from various sources undergo a semantic ETL process, resulting in CovKG. It is pivotal to ensure the correctness of the ETL process. While the concept of correctness is extensive and beyond the scope of this work, we conducted a partial assessment for CovKG by devising queries for which we already knew the answers. These queries fall into two categories: (1) Common knowledge: Globally recognized information. (2) Special knowledge: Information available from specific sources. Table 10 presents the assessment of CovKG's correctness, and the query details can be found at https://bike-csecu.com/datasets/CovKG/query.

We collected the correct answers from reliable sources. For instance, information on the number of confirmed Covid-19 cases among non-Hispanic whites in the USA and confirmation data for Ireland's Mayo County were sourced from the CDC's COVID Data Tracker (Centers for Disease Control and Prevention, 2023) and Geohive OpenData repository (*GeoHive Open Data*, 2023), respectively. Details on confirmation and death statistics for the municipality of Buren in the Netherlands were gathered from the NL Covid-19 Geo Hub repository (*NL COVID-19 Hub*, 2023). Upon reviewing Table 10, we noted that CovKG provided correct answers for all queries.

7.3. Analytical findings

We employ statistical methods to analyze CovKG, uncovering valuable insights from the MD data it represents. Our focus is on subsets of data that are more complete and contextually relevant, using localized statistics to extract insights from smaller, denser sections of the dataset. Furthermore, CovKG's linking capabilities support federated analysis, allowing insights to be derived without the need to fully integrate all information into the KG. Below, we showcase a few examples of new insights obtained across environmental, health, and socioeconomic aspects.

7.3.1. Environment aspect

Figs. 6–8 highlights the environmental aspect. Several research studies have concluded that nitrogen dioxide pollution is positively correlated with the transmission and mortality of Covid-19 (Liang et al., 2020; Lipsitt et al., 2021; Yao et al., 2021). These studies were conducted in China and the USA. In our CovKG, we integrate daily subnational air pollution data for South Africa and India. The environmental aspect can be observed through the total and average statistics of confirmed and death cases in relation to levels of nitrogen dioxide in South Africa and India, as depicted in Fig. 6. It can be seen that both the average and total of both deaths and confirmed cases are high for unhealthy levels of nitrogen dioxide. Hazardous and very unhealthy levels exist that are above unhealthy. Yet, unhealthy falls in the excessive side of the nitrogen dioxide spectrum. This observation underscores the evident positive correlation with Covid-19's epidemiology.

Figs. 7 and 8 show Bangladesh's division-wise confirmed case data in various seasons in with respect to humidity and temperature types. The seasons corresponding to summer, monsoon, and winter are respectively March–June, July–October, and November–February as per https://www.weatheronline.co.uk/reports/climate/Bangladesh.htm. It can be seen from Fig. 7 that most confirmed cases are in the very wet air and wet air cases, whereas dry or moderately wet air, i.e., less humid weathers have less confirmed cases. Northern divisions such as Sylhet, Rangpur, Rajshahi, and Mymensingh have less confirmed cases in winter during dry periods. However, as humidity increases, so does confirmed cases.

Fig. 8 illustrates the temperature categories warm-to-hot for Bangladesh, based on the weather classification by Piotrowicz and Ciaranek (2020). The temperature ranges are defined as follows: warm corresponds to 10.1–15 °C, very warm to 15.1–25 °C, and hot to

Table 10
Assessment of correctness of the ETL process in generating CovKG.

Correctness query	Type	Correct answer	CovKG's answer
How many continents are there?	Common knowledge	7	7
Which year among 2020–2022 is a leap year?	Common knowledge	2020	2020
How many occupations are there under ISCO-08 submajor group?	Common knowledge	43	43
How many nonhispanic white people died of Covid-19 in New Mexico, the USA in the month of January, 2021?	Special knowledge	180	180
How many confirmed cases were reported in Ireland's Mayo county in July 12, 2020?	Special knowledge	1,505	1,505
How many confirmed and death cases were reported in Netherland's Buren county on September 21, 2021?	Special knowledge	Confirmed: 2,807 Death : 20	Confirmed: 2,807 Death : 20

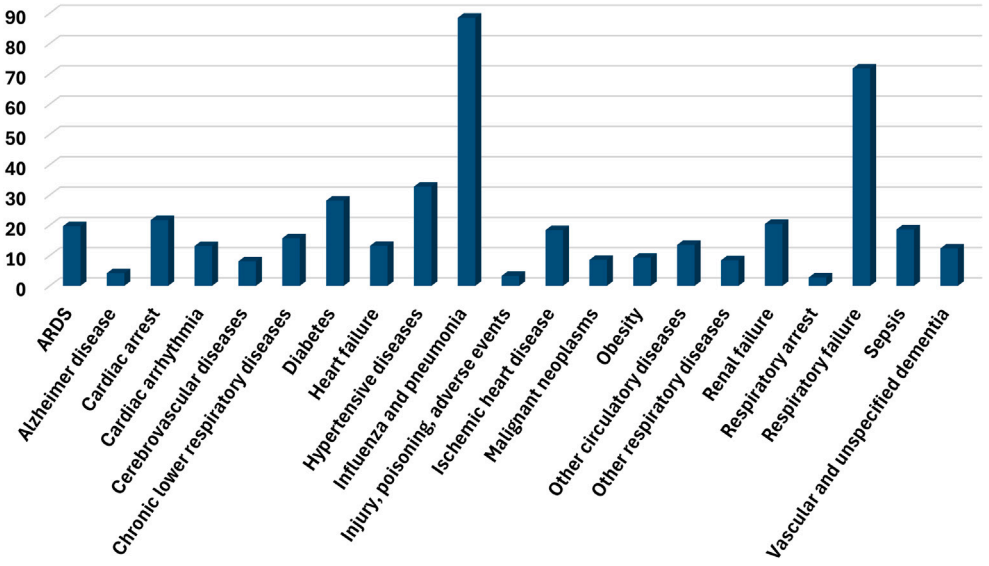


Fig. 9. Number of Covid-19 deaths per month per Admin1 unit among people with various comorbidities.

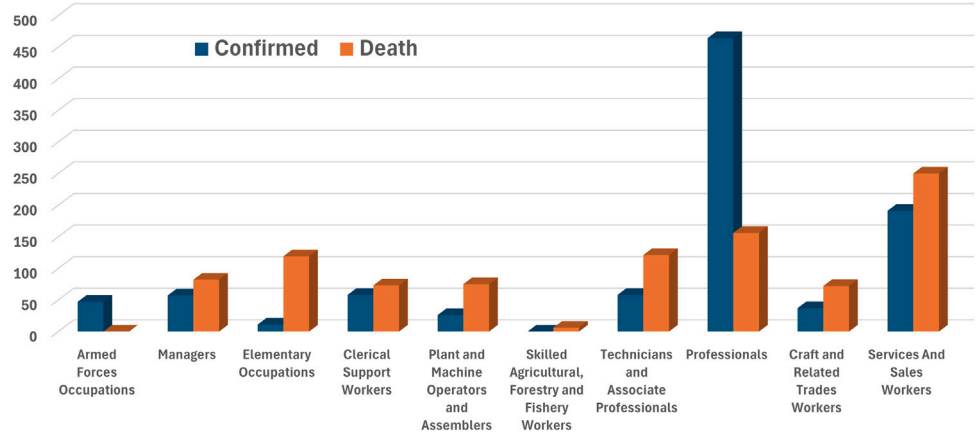


Fig. 10. Total number of deaths and confirmed cases in 2020 among ISCO-08 major occupations.

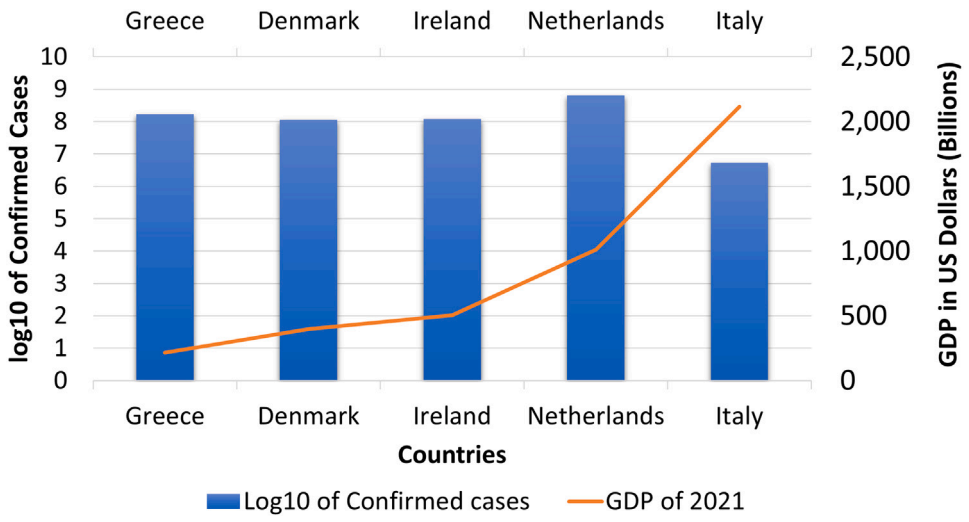


Fig. 11. The Logarithm of confirmed cases of some countries from 2021 visualized in light of their GDP in USA Dollars in that year.

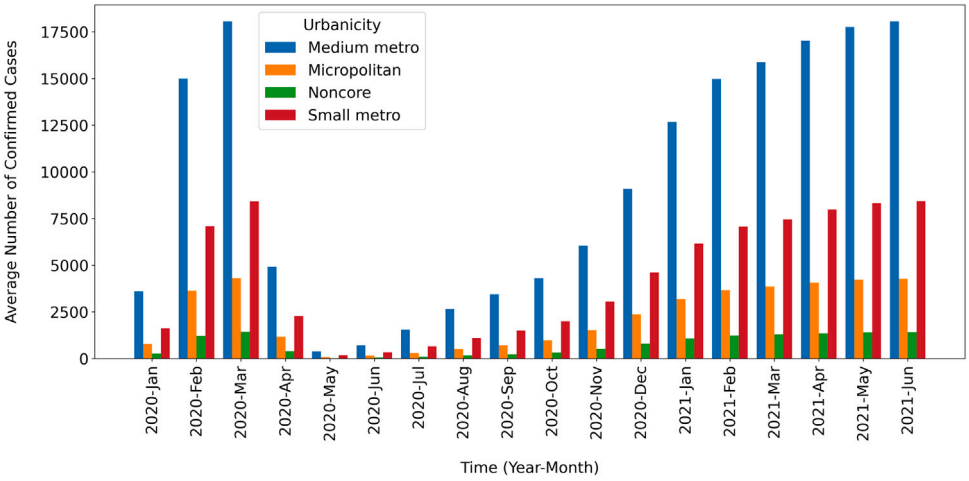


Fig. 12. Number of confirmed cases in USA areas of various urbanicity categories in the months of 2020 and 2021.

temperatures exceeding 25 °C. The data reveals that summer and monsoon seasons, characterized by higher temperatures, recorded the most confirmed Covid-19 cases, while the warm category had significantly fewer cases. Contrary to the common belief that Covid-19 spreads more in winter, the findings indicate that in Bangladesh, the virus saw the highest infection rates in warm and humid conditions. Additionally, the figures highlight that densely populated cities like Dhaka and Chittagong experienced the highest case counts during these conditions, while recording much fewer cases in less humid, colder weather. This suggests a potential positive correlation between population density, humidity, and temperature with the spread of Covid-19.

7.3.2. Health aspect

Health-related insights can be gained by examining comorbidity, as illustrated in Fig. 9. Among various comorbidities, influenza, pneumonia, and respiratory failure stand out with higher numbers of Covid-19 deaths. This correlation is noteworthy because Covid-19, influenza, and pneumonia are all respiratory diseases, suggesting that damage to the respiratory system by one of these conditions may facilitate the spread of the others. This finding aligns with previous research on the correlation between influenza and Covid-19, as shown in Alosaimi et al. (2021). Following respiratory diseases, the next highest number of deaths is observed among patients with hypertensive diseases, which are associated with high blood pressure.

7.3.3. Socioeconomic aspect

Figs. 10–13 provide insights into the socioeconomic aspect of Covid-19, focusing on the occupation, Gross Domestic Product (GDP), urbanicity, and vaccine hesitancy factors. Fig. 10 illustrates the number of confirmed and death cases among the ten major ISCO-08 occupation classes. Professionals have the highest number of confirmed cases, while services and sales workers have the highest number of deaths. This finding is intriguing because the professionals category includes medical professions such as doctors and nurses. In contrast, services and sales workers encompass professions like travel attendants, transport conductors, travel guides, waiters, cleaning and housekeeping supervisors in offices, hotels, and other establishments, as well as undertakers and embalmers. These occupations involve regular public contact. The relatively high number of confirmed cases but lower death rate among professionals suggests a level of health awareness, even in roles that require frequent public interaction.

From Fig. 11 we get a visualization of confirmed cases with respect to GDP in USA Dollar in 2021. Note that GDP data is obtained from Wikidata using a federated query. Logarithm of Base 10 was used on confirmed cases to scale down the data for better visualization. The blue bars represent the confirmed cases whereas the orange line represents the GDP. GDP is a worldwide recognized benchmark of economic condition of a country. It can be seen that there is a slight downward trend in the confirmed cases as the GDP increases. This

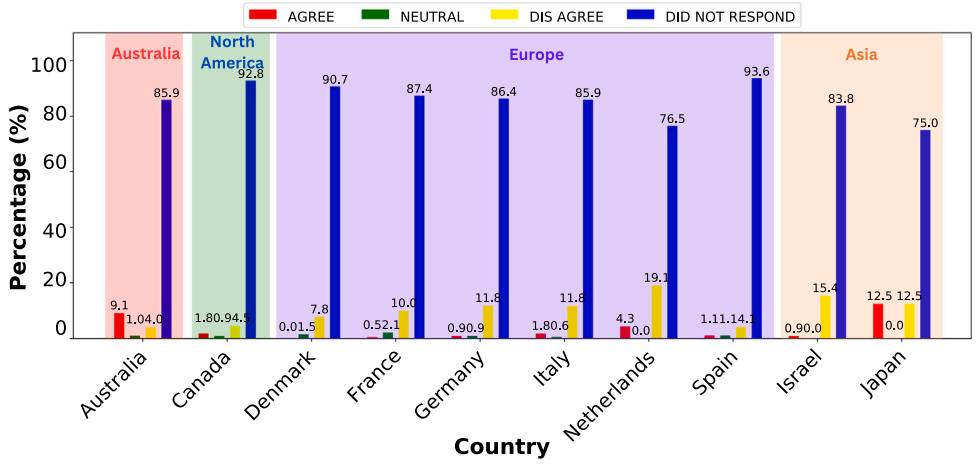


Fig. 13. Distribution of responses to the question: "If a Covid-19 vaccine were made available to me, I would definitely get it", considering only participants who had previously contracted Covid-19.

indicates an inverse correlation between GDP and spread of Covid-19. Fig. 12 shows the average confirmed case of the USA areas of various urbanicity types in the months of 2020 and 2021. It can be seen that medium metro areas consistently have the highest average confirmed cases whereas noncore areas which are predominantly rural have the lowest. This is an indication that urban areas tend to have greater chance of Covid-19 infection. Fig. 13 illustrates the distribution of responses from participants who had previously tested positive for Covid-19, regarding their willingness to accept Covid-19 vaccines, broken down by country. It is observed that only individuals from Australia and Japan, who were affected by Covid-19, are willing to receive the vaccine. In contrast, people from Europe are generally reluctant to be vaccinated after recovering from Covid-19.

8. Conclusion and future work

In this study, we generated a multidimensional and semantically annotated Covid-19 KG titled CovKG that integrates data on Covid-19 epidemiology from disparate sources and facilitates analysis from spatiotemporal, socioeconomic, health, and environmental perspectives. To our knowledge, no previous research generated a multidimensional KG dedicated to Covid-19 to such an extent as this study. CovKG allows OLAP operations and SPARQL queries to draw new insights from available data. The ETL workflow typically takes around 42 min to load CovKG, which is connected to 10,951 external resources, has a size of about 4.7 GB, and consists of about 44 million RDF triples. Moreover, due to being structured as per linked data standards, it is published as per FAIR principles, which is highly essential in cases of global phenomena like Covid-19. The qualitative assessment shows that CovKG is OLAP-compatible, can answer different aspect queries, and yields correct results when compared to other repositories. CovKG was also explored using statistical analysis to get insights into the Covid-19 situation.

CovKG faces limitations due to data sparsity, as it relies on free sources. Incorporating paid data sources and mitigating data sparsity through text conceptualization (Rahman, Nadal, Romero, & Sacharidis, 2024) could significantly improve the model, and future efforts will prioritize this improvement. Domain experts, including epidemiologists, statisticians, and WHO professionals, are better positioned to evaluate CovKG's strengths and weaknesses. To ensure a thorough assessment, we plan to collaborate with these experts. Furthermore, as a future direction, we aim to go beyond statistical analysis with SPARQL queries by enabling inter-cuboid analysis and exploring Graph Neural Network models to analyze the graph for applications such as event recognition and prediction. We also aim to explore Large Language Models with a Retrieval Augmented Generation interface to enable natural language interaction. CovKG serves as an effective use case for studying bias and fairness in KGs.

CRediT authorship contribution statement

Rudra Pratap Deb Nath: Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization, Investigation, Validation. **S.M. Shafkat Raihan:** Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization, Writing – review & editing. **Tonmoy Chandro Das:** Writing – review & editing, Conceptualization, Validation, Software. **Torben Bach Pedersen:** Writing – review & editing, Conceptualization, Formal analysis. **Debasish Ghose:** Writing – review & editing, Formal Analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research is partially supported by the Research and Publication Cell, University of Chittagong.

References

Afghanistan: Coronavirus(COVID-19) subnational cases - humanitarian data exchange. (2022). https://data.humdata.org/dataset/afghanistan-covid-19-statistics-per-province?force_layout=desktop. (Accessed on 29 August 2022).

Agapito, G., Zucco, C., & Cannataro, M. (2020). COVID-warehouse: A data warehouse of Italian COVID-19, pollution, and climate data. *International Journal of Environmental Research and Public Health*, *17*(15), 5596.

Ali, H., Alshukry, A., Marafie, S. K., AlRukhayes, M., Ali, Y., Abbas, M. B., et al. (2021). Outcomes of COVID-19: Disparities by ethnicity. *Infection, Genetics and Evolution*, *87*, Article 104639.

Alosaimi, B., Naeem, A., Hamed, M. E., Alkadi, H. S., Alanazi, T., Al Rehily, S. S., et al. (2021). Influenza co-infection associated with severity and mortality in COVID-19 patients. *Virology Journal*, *18*(1), 1–9.

Arenas, M., Bertails, A., Prud'hommeaux, E., Sequeda, J., et al. (2012). A direct mapping of relational data to RDF. *W3C Recommendation*, *27*, 1–11.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *International semantic web conference* (pp. 722–735). Springer.

Baader, F. (2003). *The description logic handbook: Theory, implementation and applications*. Cambridge University Press.

Baumgartner, R., Gatterbauer, W., & Gottlob, G. (2018). Web data extraction system.. *Encyclopedia of Database Systems, Second Edition*, *1*.

Brandt, E. B., Beck, A. F., & Mersha, T. B. (2020). Air pollution, racial disparities, and COVID-19 mortality. *Journal of Allergy and Clinical Immunology*, *146*(1), 61–63.

CCR. (2023). https://app.cpcbccr.com/ccr/#/login. (Accessed on 24 January 2023).

Centers for Disease Control and Prevention (2023). COVID data tracker.. Atlanta, GA: US Department of Health and Human Services, CDC; 2023, January 22, URL https://covid.cdc.gov/covid-data-tracker/#datatracker-home. (Accessed on 17 January 2023).

Chakrabarti, K., Chaudhuri, S., Chen, Z., Ganjam, K., He, Y., & Redmond, W. (2016). Data services leveraging bing's data assets. *IEEE Data Engineering Bulletin*, *39*(3), 15–28.

Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. *ACM Sigmod Record*, *26*(1), 65–74.

Chen, Y., Leung, C. K., Shang, S., & Wen, Q. (2020). Temporal data analytics on COVID-19 data with ubiquitous computing. In *2020 IEEE intl conf on parallel & distributed processing with applications, big data & cloud computing, sustainable computing & communications, social computing & networking (ISPA/bDCloud/socialCom/sustainCom)* (pp. 958–965). IEEE.

Chin-Hong, P., Alexander, K. M., Haynes, N., & Albert, M. A. (2020). Pulling at the heart: COVID-19, race/ethnicity and ongoing disparities. *Nature Reviews Cardiology*, *17*(9), 533–535.

Classifying the standard occupational classification 2020 (SOC 2020) to the international standard classification of occupations (ISCO-08) - office for national statistics. (2022). URL https://t.ly/OhWyx. (Accessed on 11 April 2022).

Conditions contributing to COVID-19 deaths, by state and age, provisional 2020–2022 | data | centers for disease control and prevention. (2022). https://data.cdc.gov/NCHS/Conditions-Contributing-to-COVID-19-Deaths-by-Stat/hk9y-quqm/data. (Accessed on 04 September 2022).

Data on the weekly subnational 14-day notification rate of new COVID-19 cases. (2022). https://www.ecdc.europa.eu/en/publications-data/weekly-subnational-14-day-notification-rate-covid-19. (Accessed on 29 August 2022).

COVID-19 in India | kaggle. (2022). https://www.kaggle.com/datasets/sudalairajkumar/covid19-in-india. (Accessed on 29 August 2022).

COVID, DGHS (2021). Dynamic dashboard for Bangladesh. [visited: 2021 Mar 25], URL http://dashboard.dghs.gov.bd/webportal/pages/covid19.php.

Covid19-data-greece/data/greece/regional at master · Covid-19-response-Greece/covid19-data-greece. (2023). https://github.com/Covid-19-Response-Greece/covid19-data-greece/tree/master/data/greece/regional(Accessed on 24 January 2023).

Dalvi, N., Kumar, R., & Soliman, M. (2011). Automatic wrappers for large scale web extraction. arXiv preprint arXiv:1103.2406.

Data - COVID-19 - eurostat. (2023). https://ec.europa.eu/eurostat/web/covid-19/data. (Accessed on 17 January 2023).

Data access - urban rural classification scheme for counties. (2023). https://www.cdc.gov/nchs/data_access/urban_rural.htm. (Accessed on 24 January 2023).

De Smedt, J., le Vrang, M., & Papantoniou, A. (2015). ESCO: Towards a semantic web for the European labor market. Ldow@ www.

Deb Nath, R. P., Hose, K., Pedersen, T. B., Romero, O., & Bhattacharjee, A. (2020). SETLBI: An integrated platform for semantic business intelligence. In *Companion proceedings of the web conference 2020* (pp. 167–171).

Deb Nath, R. P., Romero, O., Pedersen, T. B., & Hose, K. (2022). High-level ETL for semantic data warehouses. *Semantic Web*, *13*(1), 85–132.

Duda, O., Pasichnyk, V., Kunanets, N., Antonii, R., & Matsiuk, O. (2020). Multidimensional representation of COVID-19 data using OLAP information technology. In *2020 IEEE 15th international conference on computer sciences and information technologies, vol. 2* (pp. 277–280). IEEE.

Erling, O. (2012). Virtuoso, a hybrid RDBMS/Graph column store.. *IEEE Data Eng. Bull.*, *35*(1), 3–8.

Etcheverry, L., Gomez, S. S., & Vaisman, A. (2015). Modeling and querying data cubes on the semantic web. arXiv preprint arXiv:1512.06080.

Etcheverry, L., & Vaisman, A. A. (2012). QB4olap: a new vocabulary for OLAP cubes on the semantic web. In *Proceedings of the third international conference on consuming linked data, vol. 905* (pp. 27–38). CEUR-WS. org.

FAIR principles - GO FAIR. (2024). https://www.go-fair.org/fair-principles/. (Accessed on 3 February 2024).

Gadekar, M. C. S. (2022). Air quality index (AQI) basics. *Journal Homepage: Www. Ijrpr. Com ISSN*, *2582*, 7421.

GeoHive Open Data. (2023). https://opendata-geohive.hub.arcgis.com/. (Accessed on 24 January 2023).

GeoNames, G. (2004). The GeoNames geographical database.

Guan, W.-j., Liang, W.-h., Zhao, Y., Liang, H.-r., Chen, Z.-s., Li, Y.-m., et al. (2020). Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis. *European Respiratory Journal*, *55*(5).

Gwaze, P., & Mashele, S. H. (2018). South African air quality information system (SAAQIS) mobile application tool: Bringing real time state of air quality to South Africans. *Clean Air Journal*, *28*(1), 3.

Hâncean, M.-G., Lerner, J., Perc, M., Oană, I., Bunaciu, D.-A., Stoica, A. A., et al. (2022). Occupations and their impact on the spreading of COVID-19 in urban communities. *Scientific Reports*, *12*(1), 1–12.

Haratian, A., Fazelinia, H., Maleki, Z., Ramazi, P., Wang, H., Lewis, M. A., et al. (2021). Dataset of COVID-19 outbreak and potential predictive features in the USA. *Data in Brief*, *38*, Article 107360.

Ingram, D. D., & Franco, S. J. (2014). 2013 NCHS urban-rural classification scheme for counties. (2014), US Department of Health and Human Services, Centers for Disease Control and ....

Inmon, W. H. (2005). *Building the data warehouse*. John wiley & sons.

International Labour Office (2012). *International standard classification of occupations 2008 (ISCO-08): Structure, group definitions and correspondence tables*. International Labour Office.

Jensen, C. S., Pedersen, T. B., & Thomsen, C. (2010). Multidimensional databases and data warehousing. *Synthesis Lectures on Data Management*, *2*(1), 1–111.

Jones, S. P. (2020). Imperial college London big data analytical unit and yougov plc. 2020. Imperial College London YouGov Covid Data Hub, V1. 0, YouGov Plc.

Khan, H., Dabla-Norris, M. E., Lima, F., & Sollaci, A. (2021). *Who doesn't want to be vaccinated? Determinants of vaccine hesitancy during COVID-19*. International Monetary Fund.

Kimball, R., & Ross, M. (2011). *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons.

Knap, T., Hanečák, P., Klímek, J., Mader, C., Nečaskỳ, M., Van Nuffelen, B., et al. (2018). UnifiedViews: an ETL tool for RDF data management. *Semantic Web*, *9*(5), 661–676.

Krech, D. (2006). Rdflib: A python library for working with rdf. Online https://Github.Com/RDFLib/Rdflib.

Leung, C. K., Chen, Y., Hoi, C. S., Shang, S., & Cuzzocrea, A. (2020). Machine learning and OLAP on big COVID-19 data. In *2020 IEEE international conference on big data (big data)* (pp. 5118–5127). IEEE.

Leung, C. K., Chen, Y., Shang, S., & Deng, D. (2020). Big data science on COVID-19 data. In *2020 IEEE 14th international conference on big data science and engineering* (pp. 14–21). IEEE.

Liang, D., Shi, L., Zhao, J., Liu, P., Sarnat, J. A., Gao, S., et al. (2020). Urban air pollution may enhance COVID-19 case-fatality and mortality rates in the United States. *The Innovation*, *1*(3), Article 100047.

Lipsitt, J., Chan-Golston, A. M., Liu, J., Su, J., Zhu, Y., & Jerrett, M. (2021). Spatial analysis of COVID-19 and traffic-related air pollution in los angeles. *Environment International*, *153*, Article 106531.

Lohmann, S., Link, V., Marbach, E., & Negru, S. (2015). Webvowl: Web-based visualization of ontologies. In *Knowledge engineering and knowledge management: EKAW 2014 satellite events, VISUAL, EKM1, and ARCOE-logic, linköping, Sweden, November 24-28, 2014. revised selected papers. 19* (pp. 154–158). Springer.

McBride, B. (2004). The resource description framework (RDF) and its vocabulary description language RDFS. In *Handbook on ontologies* (pp. 51–65). Springer.

McGuinness, D. L., Van Harmelen, F., et al. (2004). OWL web ontology language overview. *W3C Recommendation*, *10*(10), 2004.

Musen, M. A. (2015). The protégé project: a look back and a look forward. *AI Matters*, *1*(4), 4–12.

Naqvi, A. (2021). COVID-19 European regional tracker. *Scientific Data*, *8*(1), 1–14.

Nath, R. P. (2020). *Aspects of semantic ETL*. Aalborg Universitetsforlag.

Nath, R. P. D., Das, T. R., Das, T. C., & Raihan, S. S. (2024). Knowledge graph generation and enabling multidimensional analytics on Bangladesh agricultural data. *IEEE Access*.

Nath, R. P. D., Hose, K., Pedersen, T. B., & Romero, O. (2017). SETL: A programmable semantic extract-transform-load framework for semantic data warehouses. *Information Systems*, *68*, 17–43.

Nath, R. P. D., Seddiqui, M. H., & Aono, M. (2014). An efficient and scalable approach for ontology instance matching.. *Journal of Computers*, *9*(8), 1755–1768.

Negash, S. (2004). Business intelligence. *Communications of the Association for Information Systems*, *13*(1), 15.

NL COVID-19 Hub. (2023). https://nlcovid-19-esrinl-content.hub.arcgis.com/. (Accessed on 24 January 2023).

Piotrowicz, K., & Ciaranek, D. (2020). A selection of weather type classification systems and examples of their application. *Theoretical and Applied Climatology*, *140*, 719–730.

Piotrowicz, K., Ciaranek, D., Wypych, A., Razsi, A., & Mika, J. (2013). Local weather classifications for environmental applications. Aerul Şi Apa. Componente Ale Mediului=Air and Water. Components of the Environment, Vol. 2013.

Rahman, M. A., Nadal, S., Romero, O., & Sacharidis, D. (2024). Mitigating data sparsity in integrated data through text conceptualization. In *2024 IEEE 40th international conference on data engineering* (pp. 3490–3504). IEEE.

Raihan, S. S., Nath, R. P. D., & Das, T. C. (2023). Covid-19 knowledge graph generation and enabling analysis across healthcare, socioeconomic, and environmental dimensions. In *2023 26th International conference on computer and information technology* (pp. 1–6). IEEE.

Roh, Y., Heo, G., & Whang, S. E. (2019). A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), 1328–1347.

Sakor, A., Jozashoori, S., Niazmand, E., Rivas, A., Bougiatiotis, K., Aisopos, F., et al. (2023). Knowledge4COVID-19: A semantic-based approach for constructing a COVID-19 related knowledge graph from various sources and analyzing treatments' toxicities. *Journal of Web Semantics*, *75*, Article 100760.

Sanyaolu, A., Okorie, C., Marinkovic, A., Patidar, R., Younis, K., Desai, P., et al. (2020). Comorbidity and its impact on patients with COVID-19. *SN Comprehensive Clinical Medicine*, *2*(8), 1069–1076.

Sari, E., Kağan, G., Karakuş, B. Ş., & Özdemir, Ö. (2022). Dataset on social and psychological effects of COVID-19 pandemic in Turkey. *Scientific Data*, *9*(1), 1–7.

Shang, S., Leung, C. K., Chen, Y., & Pazdor, A. G. (2020). Spatial data science of COVID-19 data. In *2020 IEEE 22nd international conference on high performance computing and communications* (pp. 1370–1375). IEEE.

South Africa provincial breakdown | Covid-19 South Africa. (2023). https://www.covid19sa.org/provincial-breakdown. (Accessed on 24 January 2023).

Ten threats to global health in 2019. (2022). https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019. (Accessed on 04 september 2022).

Tennison, J., Cyganiak, R., & Reynolds, D. (2012). *The rdf data cube vocabulary: Tech. rep*, Technical report, W3C Working Draft 05 April, 2012. http://www. w3. org/TR ....

The linked open data cloud. (2024). https://lod-cloud.net/. (Accessed on 03 January 2024).

Townsend, M. J., Kyle, T. K., & Stanford, F. C. (2020). Outcomes of COVID-19: disparities in obesity and by ethnicity/race. *International Journal of Obesity*, *44*(9), 1807–1809.

Travaglio, M., Yu, Y., Popovic, R., Selley, L., Leal, N. S., & Martins, L. M. (2021). Links between air pollution and COVID-19 in England. *Environmental Pollution*, *268*, Article 115859.

Turki, H., Hadj Taieb, M. A., Shafee, T., Lubiana, T., Jemielniak, D., Aouicha, M. B., et al. (2022). Representing COVID-19 information in collaborative knowledge graphs: the case of wikidata. *Semantic Web*, (Preprint), 1–32.

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

Wang, B., Li, R., Lu, Z., & Huang, Y. (2020). Does comorbidity increase the risk of patients with COVID-19: evidence from meta-analysis. *Aging (Albany NY)*, *12*(7), 6049.

Welcome - humanitarian data exchange. (2022). https://data.humdata.org/. (Accessed on 29 August 2022).

WHO coronavirus (COVID-19) dashboard. (2023). URL https://covid19.who.int/. (Accessed on 17 January 2023).

Windsor-Shellard, B., & Nasir, R. (2021). Coronavirus (COVID-19) related deaths by occupation, England and Wales: deaths registered between 9 march and 28 december.

World Bank (2023). COVID-19 household monitoring dashboard. URL https://www.worldbank.org/en/data/interactive/2020/11/11/covid-19-high-frequency-monitoring-dashboard. (Accessed on 22 January 2023).

World Weather Online (2016). Worldweatheronline.com. (Accessed on 29 August 2022).

World Wide Web Consortium, et al. (2012). R2RML: RDB to RDF mapping language.

Worldometer (2020). Coronavirus death toll and trends—Worldometer. URL https://www.worldometers.info/coronavirus/.

Wu, X., Nethery, R. C., Sabath, M. B., Braun, D., & Dominici, F. (2020). Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. *Science Advances*, *6*(45), eabd4049.

Yao, Y., Pan, J., Liu, Z., Meng, X., Wang, W., Kan, H., et al. (2021). Ambient nitrogen dioxide pollution and spreadability of COVID-19 in Chinese cities. *Ecotoxicology and Environmental Safety*, *208*, Article 111421.

Yu, L. (2011). Linked open data. In *A developer's guide to the semantic web* (pp. 409–466). Springer.