Annals of Library and Information Studies
Vol. 69, September 2022, pp. 208-220
DOI: 10.56042/alis.v69i3.58950

A survey and classification of publicly available COVID-19 datasets

Biswanath Dutta[a], Puranjani Das[b] and Sushmita Mitra[c]

[a]Associate Professor, Documentation Research and Training Centre, Indian Statistical Institute, Bangalore, Karnataka, Email: bisu@drtc.isibang.ac.in

[b]Project Linked Personnel, Documentation Research and Training Centre, Indian Statistical Institute, Bangalore, Karnataka; Email: puranjanidas02@gmail.com

[c]Professor, Machine Intelligence Unit, Indian Statistical Institute, Kolkata, West Bengal; Email: sushmita@isical.ac.in

Received: *13 April 2022*; revised and accepted: *21 July 2022*

The current study curates a list of authentic and open-access sources of alphanumeric COVID-19 pandemic data. We have gathered 74 datasets from 42 sources, including sources from 18 countries. The datasets are searched through the Kaggle and GitHub repositories besides Google, providing a representation of varieties of pandemic-related datasets. The datasets are categorized according to their sources- primary and secondary, and according to their geographical distribution. While analyzing the dataset, we came across some classes in which the datasets can be categorized. We present the categorization in the form of taxonomy and highlight the present COVID-19 data collection and use challenges. The study will help researchers and data curators in the identification and classification of pandemic data.

Keywords: COVID-19; Classification; Curation; Datasets; Metadata

Introduction

COVID-19 pandemic originated in Wuhan, China when the first case was identified in December 2019. Since then, it has spread around the globe costing 5,310,502 lives as of 15th December 2021[1]. The COVID-19 virus has changed its variants from time to time and different countries came up with several vaccines. Researchers are working day and night to tackle the various phases of the pandemic. Such phenomena and activities have contributed to the massive growth of data and datasets. The various government and non-government agencies, and individuals have been involved in data curation and building databases to increase the awareness about the disease among the common people and to support and facilitate the research in the field of COVID-19.

The researchers involved in the study have diverse backgrounds, such as, medical science, mathematics, chemical science, economics, computer science, and information science and others. The research has been undertaken in various directions and often the research is multidimensional.

For example, Ghosh et al.[2] created a discrete-time epidemic model through stability analysis of real datasets. Zoabi, Deri-Rozov, and Shomron[3] built a predictive model based on the symptoms of the affected patients. Jackson et al.[4] studied the associations between smoking and the risk of COVID-19. On the other hand, ontology-based[5] studies are also gaining prominence. Dutta and DeBellis[6] provided an ontology called CODO, a COviD-19 Ontology for data collection and analysis, utilizing the data extracted from the Government of Karnataka COVID-19 information portal[7]. CovidGraph[8], a COVID-19 knowledge graph (KG) project is aimed at supporting the researchers in finding the necessary datasets and tools. Many more such COVID-19-related studies are available in the literature. Importantly, these studies involve data and varieties of data.

The data is necessary for a deeper understanding of an underlying scenario. Only by a profound understanding of the data, new models can be built, important insights can be drawn, and a step towards further research can be taken. In pandemic times, the speedy availability of data is crucial as it allows the researchers, public health authorities, decision and policymakers, and others involved in judging and understanding the overall situation of the disease and its impact. The John Hopkins University of Medicine (JHU) was the first to release a dataset on GitHub[9].

A large volume and variety of data have already been generated in this pandemic and this is a continuous process. The amount is so huge, that there has arisen a need to organize them effectively. With the variety of data, it gets difficult to find the datasets that we need at the right time. It is necessary to analyze the datasets properly before they can be used. Also due to a huge amount of data generation, there has occurred duplication. Incorrect or incomplete data also exists. Therefore, analysis of data is strictly necessary before usage. Data analysis and organization lead to efficient and easy usage of data. Therefore, the need is to analyze the data and organize it in an efficient way. As per our knowledge, there exists no such work except for a few, e.g., Ashofteh and Bravo[10] and Shuja et al.[11]. They mainly focused on analyzing the dataset usage challenges.

The available COVID-19 datasets on the web can be broadly categorized into two: private datasets and public datasets. The private datasets refer to the datasets created and shared privately and have restricted access. Only authorized users can get access to the data, e.g., the COVID Research Database[12] requires a detailed proposal before the data access is granted. The public datasets refer to the datasets available for public use. Anyone can read and download the data. The current study focuses on public datasets.

The datasets are available in various forms, for instance, tabular data, image (e.g., chest X-ray and CT scan), audio (e.g., cough recordings COUGHVID[13]), etc. In this study, we focus on tabular data. It is usually represented in a tabular form where each column represents a particular variable, and each row represents a sample[14].

The current work is an effort toward the realization of (i) every researcher his/her data and (ii) saving the time of the researchers. The objectives of the study are (i) identification of significant sources of datasets; (ii) providing the fundamental information about them so that they become readily identifiable and selectable; (iii) domain-wise classification of the data. The other goal is to bring out the various issues regarding the COVID-19 datasets' collection, analysis, and use.

Review of literature

Alamo et al.[15] provide an analysis of the fundamental aspects of the COVID-19 domain. It suggests limitations in the various open data resources. It also mentions the country-wise facilities of open data. Ashofteh, and Bravo study the quality of COVID-19 data in three official datasets, namely the World Health Organization (WHO), European Centre for Disease Presentation and Control[16] (ECDC), and Chinese Center for Disease Control and Prevention (CCDC). A consolidated dataset was created from all three of them (available from https://data.mendeley.com/datasets/nw5m4hs3jr/2). The study presents a dataset with 11,838 rows and 37 attributes. Shuja et al. focus on the creation of a taxonomy of different kinds of open-source datasets available for COVID-19, based on datatype, applications, methods, and repositories. They covered the imaging, speech, and textual datasets. Zuo, et al.[17] conducted an analysis based on the collection of articles present at the PubMed Central on COVID-19. The study shows the distribution of datasets based on their country. It shows that 53% of them belonged to the epidemiology data. Similarly, data formats, update frequency, license, repository-wise distribution, etc. are also studied.

Cheng and Ludäscher[18] aggregated the USA COVID-19 data retrieved from the JHU dataset. The data was aggregated at the state level, county level, and regional level. The study demonstrates the analysis of case counts based on the various levels of geographic regions. Das et al.[19] review the different methods used for COVID-19 misinformation detection in existing research with an overview of data pre-processing and feature extraction methods. Szmuda et al.[20] deal with four major datasets, namely JHU, Our World in Data[21] (OWID), WHO, and ECDC. The authors gathered relevant data regarding COVID-19 to aid the researchers. The study lists potential research questions that could be tackled in the future.

Wang et al.[22] discuss the creation of the CORD dataset- Covid Open Research Dataset, which is a resource of publications related to COVID-19. The metadata and documents are collected from WHO, PubMed Central, BioRxiv, and MedRxiv. It is an ongoing task, containing around 52000 papers with 41000 full-text entries from 3200 journals. Santos et al.[23] composes a dataset of 40,212 articles on COVID-19 collected from Scopus, PubMed, arXiv, and bioRxiv from January 2019 to July 2020. The dataset can be used for natural language processing tasks. The data are tabulated according to the database they are mined from.

It can be seen that there has not been much research in the area of COVID-19 data curation and organization. Table 1 gives the differences between the related works and the current study.

Table 1 — Comparative analysis of the existing studies and the current study

Existing studies	Current study
• Existing studies are mostly related to the collection of datasets to perform a variety of statistical tasks. • Only one of the studies[11], as per our knowledge, provided a taxonomy to facilitate dataset classification by the media type (e.g., Twitter textual data, medical images, etc.), application (e.g., cough-based analysis), repositories (e.g., Github), and methods (e.g., statistical). • Neither describe the datasets nor provide the descriptors for the purpose. • Judges the quality of data in the datasets.	• Creates a detailed taxonomy with room for adding more categories in case, new varieties of data are generated during the pandemic. • Fundamentally classifies the datasets from the source of information. To further facilitate the classification, a subject-driven approach is adopted. • Provides descriptors for the datasets, which would promote the systematic and consistent dataset publication and distribution. • Evaluates the authenticity of the source and the presence of data descriptors.

Methodology

Figure 1 depicts the six steps approach that has been followed in conducting the current study. This is a generic methodology and can be applied in similar studies.

Step 1: Definition of purpose

One needs to identify and define the exact purpose of the work that must be done. Here, the purpose is the identification of datasets related to COVID-19 on the internet.

Step 2: Data repository identification

This step consists of the identification of the databases which can be searched for the defined datasets. For the current study, the Google search engine is used and later the repositories like GitHub[24] and Kaggle[25] are searched.

Step 3: Dataset search

The repositories found in the previous step are searched thoroughly. The repositories were searched using the keywords 'COVID-19', 'COVID-19 datasets', 'COVID', etc. Repositories like GitHub and Kaggle had 112,763 and 1600 search results as of 28th December 2021, respectively. We have sorted the results according to recent activity in them. The top 100 data sources from each repository are considered for the study.

While searching, we came across sources that had datasets from the government agencies and the hospitals. The other type of sources we came across curated their datasets based on the previously mentioned sources of data. Hence, the idea of

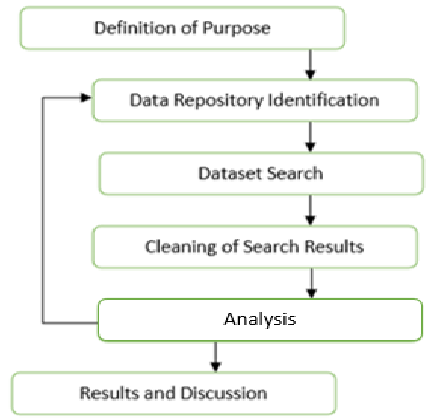

Fig. 1 — Steps for identification and analysis of COVID-19 database

segregating the datasets based on their origin as primary and secondary (aka derivational) sources arose.

Step 4: Cleaning of search results

We filter out the results available from the previous step and select the ones that best suit our purpose. The data sources that had a proper description like geographical information, chronological information, license, etc. are filtered out from the search. Also, some data sources appeared in multiple repositories among which the one with the best description is selected. For example, the *JHU* data is available on both Kaggle and GitHub, but we have listed its source as GitHub. More precisely, the datasets are selected if the following criteria are fulfilled-

- *Open Access* – the datasets can be downloaded from the source directly or by signing up free of cost on the site.
- *Listed on authentic sites* – the sites which can be referenced for academic writing.
- *Chronological information available* – the timeline of the collected data.
- *Geographical information available* – the area dealt with in the data.
- *Data description available* – description of the type of data present.

Besides these inclusion criteria, there are other factors that influenced the selection of the datasets for the current study. For example, as can be seen from Table 2, there exists a relatively small number of data sources obtained from Kaggle. It is because, despite Kaggle hosting several COVID-19-related databases, some of them do not contain datasets at all. Also,

Table 2 — Number of the data sources and datasets from GitHub and Kaggle

Repository	No. of data sources before applying the selection criteria	No. of data sources after applying the selection criteria	No. of datasets
GitHub	100	19	33
Kaggle	100	4	4
Total		23	37

some databases have restricted access. For example, covid19india-cluster[26] has a spreadsheet of data but does not provide access to everyone. COVID-19[27] appears on the first page of GitHub search results. But it is just a transformation of data from JHU into JSON format. Also, many of the COVID-19-related queries result in image-related data, which is not within the scope of the current study.

For example, the COVID-CT[28] repository, appearing on the second page of search results, contains data about CT scans. Another source, the COVID-19[29] appears on the second page of the search results and leads to python codes and not the real data. Hence, the results after being cleaned have reduced significantly. In the case of Kaggle, the very first search result that appears is the COVID-19 dataset[30] created by extracting data from JHU, which already exists in our list from GitHub. Similarly, another of the first page results, COVID-19 in India[31] is created from the information portal of the Ministry of Health and Family Welfare (MoHFW), Government of India (GOI) which exists in our list from Google search. So, there exist a lot of data sources that lead to repetitive or false results.

Following these factors, we get 23 data sources and 37 data sources together from Github and Kaggle (Table 2). We also found another 37 datasets from 19 sources. For example, ECDC, WHO, OWID, etc. are a few sources the datasets were derived from. In total, we got 74 datasets from 42 sources covering the datasets from 18 countries.

Step 5: Analysis

We analyze the datasets obtained from the previous step. We study their features and tabulate them. We also provide a taxonomy for annotating and classifying the COVID-19 datasets. This step has been detailed in the succeeding Analysis section.

Step 6: Results and discussion

The datasets are organized based on their source types: primary and secondary. Inferences gathered during the process are also mentioned. It is to be mentioned that the study does not intend to classify all the available data on the web. The study can rather be treated as a pilot study for the identification and classification of dataset sources and types.

Table 3 — Dataset features

Sl. no.	Feature Name	Feature Description
1	Name	Name of the data source from where the datasets are obtained.
2	Identifier	Data source URI.
3	Domain	Subject coverage of the dataset.
4	Geographic Information (GI)	Geographic area which the dataset covers, e.g., country, city region, continent, etc.
5	Chronological Information (CI)	Time period covered in the dataset. Some of them are ongoing datasets for which only the starting date is mentioned.
6	Frequency	Mentions the frequency with which the dataset is updated.
7	License	Licensing information of the dataset.
8	Format	File format of the dataset, e.g., CSV, JSON, etc.
9	Description	Dataset description.

Analysis

As stated above, we have gone through the datasets and studied their features. Table 3 enlists the eight features and their descriptions. These features present a clear idea about the dataset content, where to find it, and its usage. Further, Table 3 can be considered as a template by the data publishers to present new datasets. We consider that the datasets published with these minimal metadata will improve the data search and retrieval, usage, and overall, data transparency.

The dataset domains have been derived by going through the data and finding the data coverage. For example, in a dataset, from the MoHFW, GOI, the following columns were found, 'Cumulative Active Cases', 'Cumulative Cured/Discharged/Migrated', 'Cumulative Deaths', 'New Cases since yesterday', and 'New Deaths since yesterday'. We can understand that this dataset's focus is the case data. This dataset reports the latest statistics on COVID-19 cases. And this is not a single instance. Many other datasets have a similar focus, e.g., the datasets from *ECDC-Patients Data*, *OWID*, *Age Stratified COVID-19 Cases from DataPort*, and *Corona World-o-Meter (CWOM)*. This fact has led us to derive the domain name *Patient*. Similarly, a dataset of USA's state COVID-19 policies, retrieved from *GitHub* consists of data points such as 'gathering restrictions', 'school open/close', 'curfew', 'date issued', 'date ended', etc. This finding has led us to define a new domain name.

Precautionary Procedure Monitoring Status.

Continuing this process, we identified a total of twelve domains. The domains were further analyzed and grouped into their more generic categories. This has eventually led to the formulation of taxonomy with COVID-19 as its root concept (aka class) shown in Figure 2. The taxonomy is represented in two levels: array 1 representing the generic classes and array 2 representing the specific dataset domain classes. Table 4 defines the domains including their equivalent concepts from SNOMED CT[32]. The produced taxonomy is used in the current work in describing the identified datasets as provided in the result and discussion section. It is worth mentioning that as the taxonomy is built based on the existing COVID-19 datasets, we can state that the taxonomy reflects the present state of the types of available COVID-19 datasets. The taxonomy can be further extended as and when we come across new types of data.

Results and discussion

This section describes the datasets organized by their source types i.e., primary, and secondary. We have also presented a list of datasets according to their geographical locations. The descriptions are produced as per the features presented in Table 3. The datasets can be useful to the people from various domains of COVID-19 research. Most importantly, it will provide an idea regarding the types of structured data available on the pandemic.

Primary sources

Primary Sources refer to the data collected directly from the first party dealing with COVID-19, e.g., government, hospitals, Labs, etc. Table 5 describes a list of authentic COVID-19 primary datasets. For example, the first entry of the table enlists the MoHFW, GOI dataset, containing the statistical data on vaccination, patients, and testing. The dataset is in CSV format.

Secondary sources

Secondary sources refer to the data collected and curated by a second party from the primary sources. For example, the JHU dataset has been curated from the data from ECDC, Los Angeles Times, etc. A detailed list of sources is present on their GitHub page[9]. Similarly, the Indian Statistical Institute Bangalore (ISI-BC) has curated the datasets from the Indian state government's media bulletins, namely Karnataka, Maharashtra, Tamil Nadu, Andhra Pradesh, Orissa, and West Bengal. Table 6 describes a set of authentic COVID-19 secondary datasets. For example, the first entry of Table 6 enlists the JHU dataset which extracts the statistical data of patients, vaccinations, and tests from sources like ECDC (a primary source listed in Table 5), etc. The data is stored in CSV format, licensed under CC BY 4.0. The data is updated hourly.

Geographically classified sources

Table 7 describes dataset sources from various regions around the globe collected from GitHub and other data repositories. These sources provide a somewhat significant representation of all the domains of classification. This collection is basically to lead the users to a regional dataset. As it is not feasible to collect datasets of all the regional ones around the globe, we have made the collection from

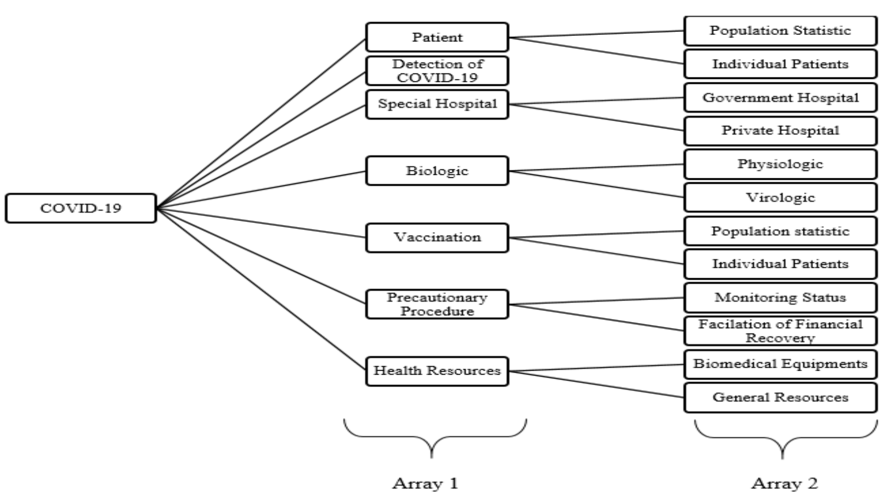

Fig. 2 — Taxonomy for classifying the COVID-19 datasets

Table 4 — Domain names and their mapping to the SNOMED CT terminologies

Levels	Domain name	Equivalent SNOMED CT term	SNOMED CT ID	Description
Array 1	Patient	Patient (person)	116154003	Datasets deal with patient (the cases) related data.
	Detection of COVID-19	Detection of Severe acute respiratory syndrome coronavirus 2 (observable entity)	871562009	Datasets deal with data related to tests conducted and positive/negative counts obtained.
	Special Hospital	Special hospital (environment)	288561005	Datasets deal with the capacity and other occupancy-related data from COVID-19 facilities.
	Biologic	Biologic (qualifier value)	12893009	Datasets deal with data generated from labs and research facilities about the virus or its symptoms.
	Vaccination	Administration of vaccine to produce active immunity (procedure)	33879002	Datasets deal with the vaccination status of the population.
	Precautionary procedure	Precautionary procedure (procedure)	389099004	Preventive measures and instructions to curb the spread of the virus.
	Health Resources	NF	NF	Available manpower, facilities, revenue, equipment, and supplies to produce requisite health care and services.
Array 2	Population Statistic	Population statistic (observable entity)	409652008	Data related to the overall population.
	Individual Patients	Individual (person)	385435006	Data related to individuals
	Government Hospital	Government hospital (environment)	79993009	COVID-19 facility aided by the Government.
	Private Hospital	Private hospital (environment)	309895006	COVID-19 facility aided by non-government agencies.
	Physiologic	Physiologic (qualifier value)	1360005	Physiological data obtained from tests conducted on COVID-19 patients.
	Virologic	Virologic (qualifier value)	7618003	Data generated from the research facilities working with the Corona Virus.
	Monitoring Status	Monitoring status (finding)	308537004	Rules and Regulations imposed to check the growth of the virus
	Facilitation of financial recovery	Facilitation of financial recovery (procedure)	711115006	Grants and donations to aid the common people in the time of the pandemic.
	Biomedical Equipment	Biomedical equipment (physical object)	303607000	Datasets consist of the necessary equipment purchased by various agencies to be supplied to the COVID-19 facilities.
	General Resources	NF	NF	Resources available for all patients to be used, for example: Beds, ICU, etc.

NF: Not Found

Table 5 — Enlists a set of primary datasets

Sl. No.	Name	Identifier	Dataset Domain	GI	CI	Frequency	License	Format	Description
1	MoHFW, GOI	https://www.mohfw.gov.in/	PPS, VPS, DC	India	Yesterday-today	Daily 08:00 IST		Tabular Data*	Official website of the GOI regarding COVID-19. It provides other necessary information about the pandemic.
2	ECDC	https://www.ecdc.europa.eu/en/covid-19/data	VPS, PPS, SH, PPMS, DC	Europe	(a) 2020 Week 53 (b)1/3/2021 (c)4/1/2020 (d)16/3/2020 (e)2020 Week15	Weekly Thursday	CC BY 4.0	XLS, CSV, JSON, XML	ECDC has 500 teams working to collect data from 196 countries.

(*Contd*.)

Table 5 — Enlists a set of primary datasets (*Contd.*)

Sl. No.	Name	Identifier	Dataset Domain	GI	CI	Frequency	License	Format	Description
3	Public Health Infobase - Data on COVID-19 in Canada	https://open.canada.ca/data/en/dataset/261c32ab-4cfd-4f81-9dea-7b64065690dc	PPS, DC	Canada	31/01/2020	Daily	Open Government License Canada	CSV – French, English	The data is provided and managed by the Health Promotion and Chronic Disease Prevention Branch (HPCDPB) of the Public Health Agency of Canada (PHAC).
4	COVID-19 Case Surveillance Public Use Data	https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf/data	PI	USA	01/01/2020	Monthly	Public Domain ©US Government	RDF, CSV, XLSX, JSON	This dataset has 27 lakh samples with 12 features. It has a unique feature called 'Race/Ethnicity. There are datasets based on other features like age, and sex.

*available as pdf and can be downloaded and converted to CSV.
PPS: Patient-Population Statistic; **PI**: Patients-Individual; **VPS**: Vaccination-Population statistic; **VI**: Vaccination-individual; **PPMS**: Precautionary Procedure-Monitoring status; **DC**: Detection of COVID-19; **PPFR**: Precautionary Procedure facilitation of financial recovery; **SH**: Special Hospital; **BV**: Virologic

Table 6 — Enlists a set of secondary datasets

Sl. No	Name	Identifier	Domain	GI	CI	Frequency	License	Format	Description
1	JHU	https://coronavirus.jhu.edu/	PPS, VPS, DC	World, USA	January 2020	Hourly	CC BY 4.0	CSV	Data is collected from a variety of sources, as listed on their GitHub page[15]. It provides visualizations and comparative analysis of countries around the globe.
2	ISI	https://www.isibang.ac.in/~athreya/incovid19/	PI, SH, VI	Indian States- Karnataka, Maharashtra, Orissa, Andhra Pradesh, Tamil Nadu, West Bengal	March 2020	Updated frequently	GNU GPLv3	CSV	This dataset allows 'contact tracing' where one can figure out the route by which the virus spreads. This dataset is curated from bulletins published by the Government of Karnataka and other media publication[7].
3	CWOM	https://data.world/aulku/coronaworldometer	PPS, DC	World	NA	Every 10 minutes	CC BY NC SA	CSV	Data is extracted from the ECDC and three additional derived features are added to it. Data visualization available.
4	Global. Health Data Science Initiative*	https://global.health/	PI	World	April 2020	Daily 12:00 UTC	CC BY 4.0	CSV	Data collected from sources like Brazil_Paraiba, Government of Colombia, Covid19CubaData, etc. Visualization of the concentration of COVID-19 positive patients in various regions available[37].
5	Global Pandemic Real Time Report	https://ncov.dxy.cn/ncovh5/view/en_pneumonia?from=dxy&source=&link=&share=	PPS	World	NA	Daily		Excel files can be saved from the webpage.	Data is collected from WHO, CDC, and media reports. Data visualization is available.
6	Neherlabs	https://github.com/neherlab/covid19_scenarios/	PPS, SH	World	NA	Different for different countries	CC BY SA 4.0	TSV	Various sources from which this data is curated is in the given link with their respective license[38].
7	IEEE Datapor*	https://ieee-dataport.org/documents/covid-19-data	PPS, BV	Variable	NA	NA	NA	CSV, JSON, PKL	It is a collection of COVID-19 datasets related to the domains of AI, medicine, virology, etc.

(*Contd.*)

Table 6 — Enlists a set of secondary datasets (*Contd*.)

Sl. No	Name	Identifier	Domain	GI	CI	Frequency	License	Format	Description
8	Virginia Hospital	https://github.com/ddenenberg71/Virginia-COVID-19-Hospital-Metrics	SH, VPS	Virginia, USA	a)07/04/2020 b)January 2021	Daily	No legal affiliation	CSV	Data gleaned by David Denenberg, from the Virginia Hospital & Healthcare Association "COVID-19 in Virginia Hospitals" dashboard[39].
9	School Closures	https://data.humdata.org/dataset/global-school-closures-covid19	PPMS	World	February 2020	Daily	CC BY	XLSX	This is a live dataset provided by UNESCO and data curation is done from 'Education: From disruption to recovery'[40].
10	Travel Restrictions	https://data.humdata.org/dataset/covid-19-global-travel-restrictions-and-airline-information	PPMS (Airline), PPMS (Travel)	World	a)July 2021 - August 2021 b) February 2020	Updatiofrequency unknown.	CC BY	CSV	This dataset is contributed by World Food Program (WFP).
11	OxCGRT	https://github.com/OxCGRT/	PPMS	World	March 2020	Updated regularly	CC BY 4.0	CSV	Systematic dataset of COVID-19 policy, from Oxford University. Project from the Blavatnik School of Governmen[41].
12	Clinical Trials	https://www.kaggle.com/parulpandey/covid19-clinical-trials-dataset	BP	World	November 2007	Updated as and when required	(DbCL) v1.0	CSV	The source consists of XML files named after their NCT numbers, an unique identifier in Clinical Trials repository.
13	Adverse Effect of Vaccine	https://www.kaggle.com/landfallmotto/covid19-vaccine-adverse-reactions-vaers-dataset	BP (Vaccination)	USA	01/01/2021-01/10/2021		CC0	CSV	Data is collected from VAERS website and preprocessed. It has Pfizer/BioNTech, Moderna, and Johnson & Johnson vaccines.
14	WHO	https://covid19.who.int/table	PPS, DC, VPS	World	a)30/12/2019	Daily 23:59 CET	CC BY NC SA 3.0 IGO	CSV	WHO collected the numbers of cases and deaths through official communications under IHR, 200, from the official ministries of health websites and social media accounts.
15	OWID	https://ourworldindata.org/coronavirus https://github.com/owid/	PPS, VPS, DC, SH	World	01/01/2020	Daily	CC BY 4.0	CSV, JSON, XLSX	There are 207 country profiles to explore, visualize and interpret the data, collected from the JHU dataset.

*One needs to sign up to access the datasets.

Table 7 — Enlists a set of geographically classified databases

Sl. no.	Country		Identifier	Domain	CI	Frequency	License	Format	Description
1	China		https://data.world/covid-19-data/china-covid-19-cases	PPS	15/01/2020-16/08/2020		CC-0	CSV	Chinese provincial and city level base maps are available in Shapefile format.
2	Japan	Tokyo	https://github.com/tokyo-metropolitan-gov/covid19/	PPS, VPS	24/01/2020	Daily	MIT	JSON	Data visualization available[42].
		National	https://github.com/kaz-ogiwara/covid19/	PPS, DC	04/02/2020-31/01/2021		NA	CSV	Data from the Ministry of Health, Labor and Welfare (MHLW).

(*Contd*.)

Table 7 — Enlists a set of geographically classified databases (*Contd.*)

Sl. no.	Country	Identifier	Domain	CI	Frequency	License	Format	Description
3	Chile	https://github.com/MinCiencia/Datos-COVID19	PPS, VPS, DC, SH	NA	NA	Special license*	CSV	Data Table led by the Ministry of Science, Technology, Knowledge, and Innovation is to use for scientific and clinical research.
4	Indonesia	https://www.kaggle.com/hendratno/covid19-indonesia	PPS	08/01/2020 - 09/07/2021		CC BY-NC-SA 4.0	CSV	Data curated from covid19.go.id, kemendagri.go.id, bps.go.id, and bnpb-inacovid19.hub.arcgis.com
5	Italy	(a)https://github.com/pcm-dpc/COVID-19/tree/master/dati-andamento-nazionale (b) https://github.com/italia/covid19-opendata-vaccini/tree/master/dati	PPS, Vaccination	(a)24/02/2020 (b) 01/01/2021	(a) Daily 18:30 IST (b) Daily	CC BY 4.0	CSV	
6	Scotland	https://github.com/bluetail14/COVID-19-in-Scotland-analysis_Dec-2020-to-June-2021	PPS, SH, VPS	December 2020- June 2021		CC BY 4.0	CSV	Data from Public Heath Scotland and National Records of Scotland.
7	Netherlands	https://github.com/Sikerdebaard/dutchcovid19data/tree/master/data	PPS, SH	NA	Hourly	NA	JSON, XLSX	Data from https://www.stichting-nice.nl/covid-19-op-de-ic.jsp and https://www.stichting-nice.nl/covid-19-op-de-zkh.jsp
8	Switzerland	https://github.com/daenuprobst/covid19-cases-switzerland	PPS, SH, DC	March 2020	Daily	NA	CSV	An interactive dashboard and a map overview is available[43,44].
9	Vietnam	https://www.kaggle.com/nhntran/vietnam-covid19-patient-dataset	PI, SH	24/01/2020-10/05/2020	NA	NA	CSV	The data is gathered by web scrapping manually from the Vietnam Ministry of Health's website[45] and mainstream media.
10	Uruguay	https://github.com/3dgiordano/covid-19-uy-vacc-data	Vaccination^{+}	February 2021	Daily	CC BY 4.0	CSV	The project was created by @3dgiordano to publicize the Uruguay's COVID-19 vaccination information. It provides updated data to OWID.
11	Norway	https://github.com/thohan88/covid19-nor-data	SH, PPS, PPFR	NA	Daily	NA	CSV	Data is updated from official sources like The Institute of Public Health and Norwegian Directorate of Health.
12	Bangladesh	https://data.humdata.org/dataset/district-wise-quarantine-for-covid-19	PPS	07/07/2020-15/12/2020	NA	NA	XLSX	Data curated by CARE Bangladesh
13	Afghanistan	https://data.humdata.org/dataset/afghanistan-covid-19-statistics-per-province	PPS	22/03/2020	Daily (Expected)	CC BY	CSV	Data is provided by the Afghanistan Ministry of Health. But after the Taliban takeover, the sincerity of data is compromised.

(*Contd.*)

Table 7 — Enlists a set of geographically classified databases (*Contd*.)

Sl. no.	Country	Identifier	Domain	CI	Frequency	License	Format	Description
14	Philippines	https://data.humdata.org/dataset/philippines-covid-19-response-who-does-what-where	PPFR	25/03/2020-06/08/2021	Not fixed	CC BY	XLSX	This data is contributed by OCHA Philippines using the 3W template (Who is Doing What and Where).
15	Germany	https://github.com/mathiasbynens/covid-19-vaccinations-germany	VPS	26/12/2020	Daily	MIT	CSV	The repository complements the OWID Our World in Data project, which includes vaccination data for Germany as a whole.
		https://github.com/dwolffram/covid19-variants/tree/main/data	PPS (Delta)	12/05/2021	Daily	NA	CSV	The repository collects data on COVID-19 variants as provided by Robert Kotch Institute, Germany for Delta Variant of COVID-19.
16	France	https://github.com/cedricguadalupe/FRANCE-COVID-19	PPS	24/01/2020 - 06/05/2020		GPL-3.0	CSV	Data collected from Regional Health Agency, Public Health France: https://www.santepubliquefrance.fr/ Geodes: https://geodes.santepubliquefrance.fr/#c=indicator&view=map2
17	USA	https://github.com/COVID19StatePolicy/SocialDistancing/blob/master/data/USstatesCov19distancingpolicy.csv	PPMS	22/03/2020	Daily (except District of Columbia discontinued from 01/08/2021	USA COVID-19 State Policy team	CSV	Dataset was created along with the article by, Christopher Adolph, Kenya Amano, Bree Bang-Jensen, Nancy Fullman, John Wilkerson[46].
		https://data.world/associatedpress/state-ppe-purchases	Biomedical Equipment	June 2020-December 2020	NA	NA	CSV	Dataset tailed more than $7 billion in purchases of PPE and high-demand medical equipment.
		https://github.com/nytimes/covid-19-data	PPS	21/01/2020	Updated daily	CC BY NC	CSV	The cases are divided in terms of cases in prisons, colleges, excess deaths, etc.

[+]Data is available with various other related information like death, region, schedule, etc.
*The license is available from https://www.minciencia.gob.cl/sites/default/files/1771596.pdf

a few of them. In Table 7, the Geographical Information (GI) is the basis of tabulation, therefore the GI is not mentioned, and we are writing that as the name of each entry. E.g., the first entry in Table 7, enlists statistical datasets of patients from China over the period of January to August 2020. The data is stored in CSV format under a CC-0 license.

It has been observed that OWID, JHU, and ECDC are the most popular data sources for statistical data regarding patients and vaccination available online. Many secondary datasets are curated by taking data from these sources. On the other hand, few of the geographical data sources listed contribute regularly to the OWID dataset. One such example is the vaccination dataset from Germany (Table 7).

The precautionary procedure datasets, containing the monitoring status, are not large in number. The most popular among them is the dataset created by OxCGRT from Oxford University (Table 6). It is interesting to note that many universities are coming up with repositories like Oxford University and JHU. Even in India, the ISI-BC (Table 6, Sl. No 2) and the Indian Institute of Technology Kanpur[33] came up with their visual representation of pandemic data.

The sources listed above can be accessed either by registering or free from the given links without any financial or academic contribution. But there are datasets and repositories which ask for a detailed research proposal to provide for the datasets, e.g., Covid Research Database and GISAID[34]. The GISAID database deals with COVID-19 lineages and variants. It employs tools to assign phylogenetic clades and lineages to genetic sequences of the virus. The site contains various interactive dashboards related to genomic variability, global testing, variant report, sequencing report, etc. For obtaining the datasets one has to submit a proposal and agree to their terms and conditions. There exist other types of repositories (e.g., Euler Registry[35]) that do not provide data but a statistical report of the data, providing an overall summary of the situation. In further continuation of the discussion, the following section reports the COVID-19 data collection and use challenges.

COVID-19 data collection and use challenges

This section summarizes the various issues and challenges that we experienced during the data collection and analysis for the current study. We believe that they will aid in better publication and collection of the data.

Variety of formats, lack of unification

Upon carefully observing the datasets, it can be seen that the features present in the dataset overlap in multiple datasets of the same domain, but due to the lack of a common term each of them has a different term to denote them. Also, some datasets have lots of features and some of them have the bare minimum. So, even if we try to merge them, it is impossible with so many inconsistencies in the data format[17].

Lack of datasets on specific areas

As mentioned, there are not many datasets in the ‘biomedical equipment’ domain. This data although not directly related to the health of the citizens but can be used to figure out the preparedness of the particular region in matters of products like sanitizers, masks, PPE, etc.

Also, the 'precautionary procedure' dataset is limited to a few countries. E.g., India does not have a dedicated precautionary procedure dataset, which would have been incredibly helpful. The response measures are so varied because each region comes up with its level of strictness and guidelines. This disparity has to be dealt with before a significant representation of the COVID-19 situation in each country is dealt with. There are tons of unstructured and scattered data in this domain on the web but unorganized data is barely of any use. Another noteworthy observation is the availability of clinical data in the public domain, which is limited and is compromised in quality.

Bias in data availability

Most of the datasets originate from Europe and/or America. There is a lack of dedicated datasets for the countries like Afghanistan, Bangladesh, etc. Though this study contains a few datasets from those countries, in the overall scenario it is rare. As observed, the data curated from some of the countries do not guarantee accuracy and is sometimes incomplete. It is known from the COVID-19 reports, that the strain of the coronavirus is different in different regions and affects the population differently. Therefore, having data from a handful of countries does not represent the worldwide scenario accurately. Also, the datasets available from many of the countries are mostly statistical—like patient counts or vaccination counts. 'Biologic' datasets like lab data or symptoms are usually unavailable. The same goes for the responses data, which are also mostly not available.

Artificially built datasets

Apart from the real-time data available from governments and various agencies, the internet is filled with datasets that are artificially created for performing various data science tasks or activities. Students are trying to build machine learning models using those datasets. It is difficult to sieve out the authentic datasets from these.

Data description unavailable

We came across several datasets that do not consist of any description of the features whatsoever. The datasets without description may be interpreted by the domain expert, but not necessarily by data scientists or analysts whose aim is to achieve insights through data-driven methods. This is particularly more prominent in the case of clinical/epidemiological data. E.g., South Africa major lineage dataset is available on GitHub[36]. One can figure out by the name that the data is related to major lineages affected by the COVID-19 in South Africa. But no description is available for the columns in the datasets.

Comparison metric inconsistency

The situation of a given region is judged based on the metrics calculated from the statistical datasets. There is no one metric used by all the regions. This leads to the judgment of the COVID-19 situations in different regions differently. This makes the situation comparison, and analysis of those situations in various regions extremely difficult[15].

Error in data collection

There exist inconsistencies in numbers in major datasets, like WHO, ECDC, and CCDC[10].

Diversified sources

The secondary sources curate data from sources like media, hospitals, and other primary sources. This might lead to data repetition and result in numbers not aligning with reality.

Most of the datasets provide the numbers of affected, deceased, cured, etc. These are albeit necessary information, but it is also a source of panic among the common public. Instead, the datasets which are fewer in numbers, e.g., the precautionary procedure, special hospital, and biomedical equipment should be focused upon. Because they will serve in educating the mass about the state's preparedness during the pandemic along with the steps, they should be taking up to flatten the curve.

Conclusion

In this study, a curated list of 74 datasets from 42 data sources is presented along with their important features. The datasets found in those sources are classified according to their origin (primary and secondary) and the type of data present in them. A taxonomy has also been developed. It can be repurposed for classifying the COVID-19 datasets. The study also detailed the present COVID-19 data collection and use challenges.

The number of datasets on the internet is numerous, hence the present study could not accommodate an exhaustive list of all those datasets. Also, the number of datasets collected for each type is not equal. This is because of the lack of availability of those types of datasets. As discussed, there is a bias in the available datasets and although there exists a handful of statistical datasets, responses and medical data are scarce.

The current study can be expanded in various directions. The taxonomy presented can be expanded to include other types of data, which are generated in the future. Also, the present study deals with alphanumeric tables only. In continuation to the current work, our immediate aim is to include the audio and visual data (e.g., cough samples and CT scans). Apart from this, based on the analysis of various types of datasets, a uniform template can be created for each type. This will not only result in the systematic collection of data but also lead to comparative analysis between datasets of the same type.

The authors are affirmative that the proposed classification, taxonomy, and data description template will significantly impact the way datasets are shared and distributed. The authors further intend to validate the proposed study with the help of the community in the future (e.g., through survey research).

Acknowledgment

This work is executed under the research project entitled "Integrated and Unified Data Model for Publication and Sharing of prolonged pandemic data as FAIR Semantic Data: COVID-19 as a case study", funded by Indian Statistical Institute Kolkata.

References

1 World Health Organization COVID-19 Dashboard, Available at https://covid19.who.int/ (Accessed on 14 Nov 2021).

2 Ghosh D, Santra P K, Mahapatra G S, Elsonbaty A, and Elsadany A A, A discrete-time epidemic model for the analysis of transmission of COVID19 based upon data of epidemiological parameters, *The European Physical Journal Special Topics*, (2022). https://doi.org/10.1140/epjs/s11734-022-00537-2.

3 Zoabi Y, Deri-Rozov S, and Shomron N, Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *Npj Digital Medicine*, 4 (1) (2021). https://doi.org/10.1038/s41746-020-00372-6.

4 Jackson S E, Brown J, Shahab L, Steptoe A, and Fancourt D, Covid-19, smoking and inequalities: A study of 53002 adults in the UK, *Tobacco Control*, 30(e2) (2020) e111-e121. https://doi.org/10.1136/tobaccocontrol-2020-055933.

5 Dutta B, Examining the interrelatedness between ontologies and Linked Data, *Library Hi Tech*, 35 (2) (2017) 312-331.

6 Dutta B and DeBellis M, CODO: an ontology for collection and analysis of COVID-19 data, *In Proceedings of the paper presented at the 12th Int. Conf. on Knowledge Engineering and Ontology Development (KEOD)*, Lisboa, Portugal, 2-4 November 2020, 2, p.76-85.

7 COVID-19 media bulletin, Available at https://covid19.karnataka.gov.in/govt_bulletin/en (Accessed on 11 April 2022)

8 Covidgraph- A covid-19 knowledge graph. HealthECCO. (November 30 2021). Available at https://healthecco.org/covidgraph/ (Accessed on 11 April 2022).

9 CSSEGISandData. (n.d.). CSSEGISANDDATA/covid-19: Novel coronavirus (COVID-19) cases, provided by JHU CSSE. GitHub. Available at https://github.com/CSSEGISandData/COVID-19 (Accessed on 11 April 2022)

10 Ashofteh A and Bravo J M, A study on the quality of novel coronavirus (COVID-19) official datasets, *Statistical Journal of the IAOS*, 36 (2) (2020) 291–301.

11 Shuja J, Alanazi E, Alasmary W, and Alashaikh A, COVID-19 open-source datasets: a comprehensive survey, *Applied Intelligence*, 51(3) (2020) 1296–1325.

12 COVID-19 research database. Available at https://covid19researchdatabase.org/ (Accessed on 11 April 2022).

13 Orlandic L, Teijeiro T, and Atienza D, The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms, *Sci Data*, 8 (2021) 156.

14 Renear A H, Sacchi S, and Wickett K M, Definitions of the dataset in the scientific and technical literature, *In Proceedings of the American Society for Information Science and Technology*, 2010 P.1–4.

15 Alamo T, Reina D, Mammarella M, and Abella A, Covid-19: Open-Data Resources for Monitoring, Modeling, and Forecasting the Epidemic. *Electronics*, 9 (5) (2020) 827.

16 ECDC, Available at https://www.ecdc.europa.eu/en/covid-19/data (Accessed on 4 May 2022).

17 Zuo X., Chen Y, Ohno-Machado L, and Xu H, How do we share data in COVID-19 research? A systematic review of COVID-19 datasets in PubMed Central Articles, *Briefings in Bioinformatics*, 22(2) (2020) 800–811

18 Cheng Y-Y and Ludäscher B, Through the magnifying glass: Exploring aggregations of COVID-19 datasets by county, state, and taxonomies of U.S. regions. In *Proceedings of the Association for Information Science and Technology*, 57(1) (2020), p. e355. https://doi.org/10.1002/pra2.355.

19 Ullah S A R, A Survey of COVID-19 Misinformation: Datasets, Detection. Preprint at arxiv:2110.00737v1 (2021).

20 Szmuda T, Ali S, Özdemir C, Syed M T, Singh A, et al, Datasets and future research suggestions concerning SARS-CoV-2. S. *European Journal of Transitional Clinical Medicine*. 3(2) (2020) 80-85.

21 Our World in Data, Available at https://ourworldindata.org/coronavirus (Accessed on 4 May 2022).

22 Wang L L, CORD-19: The Covid-19 Open Research Dataset. https://pubmed.ncbi.nlm.nih.gov/32510522/ (2020).

23 Santos B S, Silva I, Ribeiro-Dantas M D C, Alves G, Endo P T and Lima L, COVID-19: A scholarly production dataset report for research analysis. *Data in Brief*, 32 (2020) 106178.

24 GitHub, Available at https://github.com/ (Accessed on 4 May 2022).

25 Kaggle, Available at https://www.kaggle.com/ (Accessed on 4 May 2022).

26 World Health Organization, Available at https://www.who.int (Accessed on 4 May 2022).

27 Covid19, Available at https://github.com/pomber/covid19 (Accessed on 4 May 2022).

28 COVID-CT, Available at https://github.com/UCSD-AI4H/COVID-CT (Accessed on 4 May 2022).

29 Covid-19, Available at https://github.com/k-sys/covid-19 (Accessed on 4 May 2022).

30 Devakumar K P, Covid-19 dataset. Available at https://www.kaggle.com/imdevskp/corona-virus-report (Accessed on 4 May 2022).

31 Covid-19 in India, Available at https://www.kaggle.com/sudalairajkumar/covid19-in-india (Accessed on 4 May 2022).

32 International, S. N. O. M. E. D. (n.d.). SNOMED International's SNOMED CT browser. SNOMED International Browser. Available at https://www.snomedbrowser.org/ (Accessed on 11 April 2022).

33 Covid prediction. Available at https://covid19-forecast.org/ (Accessed on 11 April 2022)

34 GISAID. Available at https://www.gisaid.org/ (Accessed on 11 April 2022)

35 EULAR. Available at https://www.eular.org/eular_covid_19_registry.cfm (Accessed on 11 April 2022)

36 SARSCoV2_South_Africa_major_lineages. Available at https://github.com/krisp-kwazulu-natal/SARSCoV2_South_Africa_major_lineages/ (Accessed on 11 April 2022)

37 Global.health: A data science initiative. Available at https://data.covid-19.global.health/data-acknowledgments (Accessed on 11 April 2022).

38 Covid19_scenarios_data. Available at https://github.com/neherlab/covid19_scenarios_data (Accessed on 11 April 2022).

39 Virginia Hospital covid-19 dashboard. Communications. (15 January 2021). Available at https://www.vhha.com/communications/virginia-hospital-covid-19-data-dashboard/ (Accessed on 11 April 2022).

40 Education: From disruption to recovery. UNESCO. (28 February 2021). Available at https://en.unesco.org/themes/education-emergencies/coronavirus-school-closures (Accessed on 11 April 2022).

41 Covid-19 government response tracker. Blavatnik School of Government. (n.d.). Available at https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker (Accessed on 11 April 2022).

42 Latest infection trends in Tokyo. Available at https://stopcovid19.metro.tokyo.lg.jp/ (Accessed on 11 April 2022)

43 Covid-19 info Switzerland. COVID-19 Info. Available at https://www.corona-data.ch/ (Accessed on 11 April 2022).

44 Corona figure dump, Available at http://corona-ch.surge.sh/ (Accessed on 11 April 2022).

45 NCov. Available at https://ncov.vncdc.gov.vn/ (Accessed on 11 April 2022).

46 Adolph, C., Amano, K., Bang-Jensen, B., Fullman, N., & Wilkerson, J., Pandemic politics: Timing state-level social distancing responses to covid-19, *Journal of Health Politics, Policy and Law*, 46(2) (2021) 211–233. https:// doi.org/10.1215/03616878-8802162