

## COVID-19 Data Hub

Emanuele Guidotti<sup>1</sup> and David Ardia<sup>2</sup>

**DOI:**

1 University of Neuchâtel, Switzerland 2 HEC Montréal, Canada

**Software**

- [Review](#) ↗
- [Repository](#) ↗
- [Archive](#) ↗

**Submitted:**

**Published:**

**License**

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](#)).



### Summary

The goal of [COVID-19 Data Hub](#) is to provide the research community with a **unified** dataset helpful for a better understanding of COVID-19.

In December 2019 the first cases of pneumonia of unknown etiology were reported in Wuhan city, People’s Republic of China.<sup>1</sup> Since the outbreak of the disease, officially called COVID-19 by World Health Organization (WHO), a multitude of papers have appeared.

By one estimate, the COVID-19 literature published in January-May 2020 has reached more than 23,000 papers and is doubling every 20 days—among the biggest explosions of scientific literature ever.<sup>2</sup>

In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (Wang et al., 2020), a resource of over 134,000 scholarly articles about COVID-19, SARS-CoV-2, and related coronaviruses.

The Center for Systems Science and Engineering at the Whiting School of Engineering, with technical support from ESRI and the Johns Hopkins University Applied Physics Laboratory, is maintaining an interactive web-based dashboard to track COVID-19 in real time (Dong, Du, & Gardner, 2020). All data collected and displayed are made freely available through a GitHub repository.<sup>3</sup>

A team of over one hundred Oxford University students and staff from every part of the world is collecting information on several different common policy responses governments have taken. The data are aggregated in The Oxford COVID-19 Government Response Tracker (Hale, Webster, Petherick, Phillips, & Kira, 2020).

<sup>1</sup>World Health Organization, Novel Coronavirus (2019-nCoV) SITUATION REPORT 1 - 21 JANUARY 2020

<sup>2</sup><https://doi.org/10.1126/science.abc7839>

<sup>3</sup><https://github.com/CSSEGISandData/COVID-19>

Google and Apple released mobility reports<sup>45</sup> to help public health officials. Governments all over the world are releasing COVID-19 data to track the outbreak as it unfolds.

Yet, to our knowledge, there is no application that collects COVID-19 worldwide fine-grained data, includes demographics, environmental factors, or other exogenous variables, and harmonizes the amount of heterogeneous data that have become available.

## Data collection

We implemented an extendable R package (R Core Team, 2020) designed to aggregate the data from several sources. Hosted on GitHub,<sup>6</sup> the package allows contributors to collaborate on the implementation of additional data sources.

The data are hourly crunched by a dedicated server exploiting the package. The data are harmonized in csv format and made available on a cloud storage, in order to be easily accessible from R, Python, MATLAB, Excel, and any other software.

Out-of-the-box packages are available for seamless integration with the Data Hub. Namely, the **COVID19** R package available on CRAN and the **covid19dh**<sup>7</sup> Python package available on PyPI. According to the terms of use, World Bank, Google, and Apple data cannot be stored on external servers, but the packages provide functionalities to download such data from the original repositories and extend the dataset in real-time. The packages use an internal memory caching system so that the data are never downloaded twice. This is especially suited for interactive frameworks, such as Shiny (Chang, Cheng, Allaire, Xie, & McPherson, 2020).

We do our best to guarantee the data quality and consistency: a) all sources are properly documented, along with their citation; b) we generate error logs to spot misalignments in the official data and inform authorities; c) we rely on the open-source community: the bigger the community, the faster possible bugs will be notified and fixed. Vintage data, daily snapshots of the data, are provided so to ensure research reproducibility. Still, this is free software and comes with absolutely no warranty.

At the time of writing, the dataset includes:

- **standard COVID-19 variables:** total population, cumulative number of cases, tests, deaths, recovered, daily number of hospitalized, patients requiring ventilation and intensive therapy.
- **policy measures** by Oxford COVID-19 Government Response Tracker (Hale et al., 2020)
- **geographic information** suited for data visualization and for interfacing with external databases (e.g. weather information, geo-located tweets).
- **external identifiers** allowing to extend the dataset with World Bank Open Data, Google mobility reports, and Apple mobility reports. Governmental identifiers are provided to further extend the dataset with local, fine-grained statistics.

The data are available at different levels of granularity: 1) administrative area of top-level, usually countries; 2) states, regions, cantons; 3) cities, municipalities. Refer to the [dataset documentation](#) for more details and to the introductory video available at [COVID-19 Data Hub](#) for an overview of the project.

<sup>4</sup><https://www.google.com/covid19/mobility/>

<sup>5</sup><https://www.apple.com/covid19/mobility/>

<sup>6</sup><https://github.com/covid19datahub/COVID19/>

<sup>7</sup>we acknowledge the efforts of Martin Beneš for providing the package

## Acknowledgements

We are grateful to the Institute for Data Valorization [IVADO](#) and [HEC Montréal](#) for sponsoring the development of the data hub. We also acknowledge the efforts of all the [volunteers](#) taking part in the data collection as a joint effort against COVID-19.

## Terms of use

You assume full risk for the use of [COVID-19 Data Hub](#). We try our best to guarantee the data quality and consistency and the continuous filling of the Data Hub. However, it is free software and comes with ABSOLUTELY NO WARRANTY. Reliance on [COVID-19 Data Hub](#) for medical guidance or use of [COVID-19 Data Hub](#) in commerce is strictly prohibited. **License:** [GPL-3](#).

## References

- Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2020). *Shiny: Web application framework for R*. Retrieved from <https://CRAN.R-project.org/package=shiny>
- Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5), 533–534. doi:10.1016/S1473-3099(20)30120-1
- Hale, T., Webster, S., Petherick, A., Phillips, T., & Kira, B. (2020). Oxford covid-19 government response tracker. *Blavatnik School of Government*.
- R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., et al. (2020). COVID-19: The covid-19 open research dataset. *arXiv preprint arXiv:2004.10706*.