From collection to analysis: A comparison of GISAID and the Covid-19 Data Portal

Nathanael Sheehan, Sabina Leonelli, Federico Botta University of Exeter

Author Note

This project has received funding from the European Research Council (ERC) under the European Union's Horison 2020 research and innovation programme (grant agreement No. 101001145). This paper reflects only the author's view and that the Commission / Agency is not responsible for any use that may be made of the information it contains.

Abstract

We analyse ongoing efforts to share genomic data about SARS-COV-2 through a comparison of the characteristics of the Global Initiative on Sharing All Influenza Data and European Nucleotide Archive infrastructures with respect to the representativeness and governance of the research data therein. We focus on data and metadata on genetic sequences posted on the two infrastructures in the period between March 2020 and October 2022, thus capturing a period of acute response to the COVID-19 pandemic. Through a variety of data science methods, we compare the extent to which the two portals succeeded in attracting data submissions from different countries around the globe and look at the ways in which submission rates varied over time. We go on to analyse the structure and underlying architecture of the infrastructures, reviewing how they organise data access and use, the types of metadata and version tracking they provide. Finally, we explore usage patterns of each infrastructure based on publications that mention the data to understand how data reuse can facilitate forms of diversity between institutions, cities, countries, and funding groups. Our findings reveal disparities in representation between the two infrastructures and differing practices in data governance and architecture. We conclude that both infrastructures offer useful lessons, with GISAID demonstrating the importance of expanding data submissions and representation, while the COVID-19 data portal offers insights into how to enhance data usability.

Introduction

The pursuit of effective data sharing in scientific research is an inherently complex and multifaceted endeavour, giving rise to divergent modes of governance, management, and stewarding. Against this backdrop, this paper undertakes a critical examination of two models of data sharing that featured prominently in the context of COVID-19 genomic data. Specifically, we explore salient characteristics of the two leading data infrastructures 1for sharing COVID-19 genomic data, Global Initiative on Sharing Avian Influenza Data (GISAID and the Covid-19 Data portal (CV19-DP), which have played a pivotal role in providing comprehensive and up-to-date genetic information on the SARS-CoV-2 virus throughout the pandemic.

GISAID is a globally operating scientific organization that serves as a primary source of open access genomic data pertaining to influenza viruses. The organisation's genesis in 2008 marked a departure from the public domain sharing model that had been espoused by European bioinformatic institutes, instead embracing a distinct data governance approach that incentivises and recognises data depositors. Through collaborative analysis of sequence data, the GISAID model of data sharing represents a paradigm shift that levels the epistemic playing field between higher and lower resourced countries. In contrast, CV19-DP, a newer data infrastructure launched in April 2020 by the European Nucleotide Archive, primarily focuses on data reuse and portability in response to the COVID-19 pandemic. In addition to viral sequence data, the platform hosts a variety of other genetic data, including host sequences, gene expressions, and proteins data all of which are completely open to reuse without any tracking of access.

In this paper we compare the characteristics of the GISAID and CV19-DP infrastructures in three respects: (1) submission rates; (2) architecture and governance; and (3) data usage. Through a variety of aggregation methods, we compare the extent to which the two portals succeeded in attracting data submissions from different countries around the globe and look at the ways in which submission rates varied over time. We further analyse the structure and underlying architecture of the infrastructures, reviewing how they organise data access and use, the types of metadata and version tracking they provide, and the scope they offer for data re-use. Our analysis is grounded on data and metadata on genetic sequences posted on the two infrastructures in the period between March 2020 and October 2022, thus capturing the period of arguably most acute response to the COVID-19 pandemic. We conclude with a reflection on what these infrastructures have achieved in terms of both their representativeness and openness.

Background: A tale of two Data infrastructures

GISAID was launched in 2008, on the anniversary of the Spanish influenza, to foster the sharing of influenza genomic data securely and responsibly. Data sharing was immediately conceptualised not as straightforward 'opening up' of the data by placing them online without restrictions to access and re-use, but rather as an alternative to the public sharing model, whereby users agree to authenticate their academic identity and not to republish or link GISAID genomes without permission from the data producer. This requirement stems from the recognition that some researchers – commonly located in low-resourced environments – are reluctant to share data due to fears of better-equipped researchers building on such work without due acknowledgment (Elbe and Buckland-Merrett 2017; Bezuidenhout and Chakauya 2018). This model proved successful in relation to influenza research, and since its launch GISAID has played an essential role supporting data sharing among the WHO Collaborating Centers and National influenza Centers in response to the bi-annual influenza vaccine virus recommendations by the WHO Global influenza Surveillance and Response System (GISRS). It is no surprise therefore that GISAID was swiftly redeployed, in early 2020, to include SARS-COV-2 data through the EpiCov database, which stores, analyses and builds evolutionary trees of SARS CoV-2 genome sequences and hosts several daily updates of visualisations (Khare et al 2021). In addition to this, GISAID The Epi-CoV database provides eleven tools to explore SARS-CoV-2 sequence data, within in these include Audacity – global phylogeny of hCoV-19 as a downloadable newick tree file -, CoVizue- near real-time visulsation of SARS-CoV-2 genetic variation –, such as a global submissions and lineage tracker, as well as a collection of thirty-two analysis figures updated daily – these are documented in table 1.

GISAID is now the leading open access database for SARS-CoV-2, with over 15 million genomes sequenced by February 2023. GISAID plays a key role in identifying and studying variant evolution, lineages, and spread in real-time; and indeed, it features as key data provider for a wide variety of consortia, initiatives and projects devoted to the analysis of COVID-19 variants of interest (some of which listed on this page:

https://gisaid.org/collaborations/enabled-by-hcov-19-data-from-gisaid/). Accordingly, GISAID is funded by a wide consortium of public and private bodies, including the Federal Republic of Germany, who first backed the project at its main site in Geneva, as well as public-health and academic institutions in Argentina, Brazil, China, Republic of the Congo, Ethiopia, Indonesia, Malaysia, Russia, Senegal, Singapore, South Africa and many other countries, as well as several donors and partners garnered under the label of "Friends of GISAID".

The GISAID model has fostered trust and information exchange among groups that differ considerably in their geo-political locations, funding levels, material resources and social characteristics, thereby expanding the range of data sources shared online (Shu and McCauley 2017). At the same time, GISAID has been frequently scrutinised as limiting the extent to which data can be accessed and linked, thereby negatively affect the insight, pace and breadth of future research – leading to the backlash by hundreds of leading researchers concerned about the urgency of an effective pandemic response (ENA 2021). During the height of the pandemic, questions were raised regarding the quality and integrity of metadata coming from the GISAID platform (Gosashti and Corbett-Detig 2021), as well as epistemic importance being placed on the lag in submissions times at the global scale (Kalia et al. 2021). Some scientists called for a complete opening of genomic data sharing for SARS-CoV-2 (Van Noorden 2021), stating that GISAID's policy may be "open data", but it does not make the data easily shareable (Yehudi et al 2022). These critiques have run alongside controversy around who retains ownership of the data stored in GISAID. During past viral outbreaks, GISAID has been involved in legal disputes between the Swiss Institute of Bioinformatics (SIB) over monetary funds of the infrastructure (Greenemeier 2009), such events led to a spokeswoman at GISAID asserting that the SIB had misappropriated the database on grounds of data ownership (Butler 2009).

In April 2020, the European Commission, through the auspices of the European Molecular Biology Laboratory (EMBL) and Elixir – a twenty-three country node network in Europe dedicated to openly sharing life scienc data -, launched another platform for sharing scientific – including genomic – information of relevance to the biological study of COVID-19: the COVID-19 Data Portal (CV19-DP) (Harrison et al 2021). The data infrastructure hosts a diverse array of data types, including protein, expression, networks, imaging, and socio-economic data, encouraging data linkage and cross-analysis (Saravanan et al. 2022). By design, CV19-DP is modular, enabling swift development of customised versions of the data portal by EU nations, as exemplified by the Spanish version at https://covid19dataportal.es and the Polish version at https://covid19dataportal.pl. Nevertheless, the primary site, located at https://covid19dataportal.org, remains a key point of entry.

The CV19-DP aims to facilitate research by promoting data sharing and enhancing data interoperability with other infrastructures such as ENA, UniProt, PDBe, EMDB, Expression Atlas, and Europe PMC. To accomplish this, the CV19-DP utilises a high-level application programming interface (API) and direct bulk downloads with minimal user tracking to distribute data transparently and efficiently. The portal offers two key data visualisation tools for the rapid analysis of the large volume of data stored on the system. The first is an open-source phylogenetic tree that illustrates Covid-19 sequences that represent 98% of the SARS-CoV-2 template, including PANGO lineages, and are stratified by WHO regions. The second is the CoVeo browser, a proprietary software that performs a systematic analysis of raw reads from CV19-DP and presents the results in graphical and summary form for global regions, specifically examining VOCs and VOIs.

Both GISAID and CV19-DP ostensibly adhere to the principles of fairness, but their respective approaches reflect the nuanced ways in which these principles can be interpreted and applied in practice. In the case of CV19-DP, fair data is predicated on the FAIR data principles, which espouse the findability, accessibility, interoperability, and reusability of research data (Wilkinson et al., 2016) in order to ensure machine readability (Becket REF). While this framework is designed to promote transparency and data sharing, it remains subject to interpretation and implementation by individual data repositories, with potential variations in compliance and enforcement across different contexts (Boeckhout et al.2018). Conversely, GISAID's interpretation of fairness is best understood through its database access agreement, which seeks to enable unrestricted and prompt access to epidemic and pandemic virus data. This approach foregrounds equitable exploitation of results, scientific etiquette, and open data sharing while safeguarding intellectual property rights (paraphrased from the GISAID website:).

Even though both data infrastructures share common epistemic goals, cognitive-cultural resources, and knowledge forms on complex biological systems, they create distinct digital artifacts as a result of different policies and values. Arita (2021) points out how the data infrastructures make use of different understandings of open access – with GISAID being understood as "partially closed" and the CV19DP as "fully open". Bernasconi et al. (2021) conclude that, while GISAID's partially closed model is likely to attract international collaboration from under-resourced countries, it fails to provide features of data provenance such as persistent URLs to samples or publications. The urgency to better understand the epistemic role of these infrastructures – and those to come after it - is underscored by the work of Chen et al. (2022) and Brito et al. (2022) who identified that countries in lower income groups often lack efficient genomic surveillance capabilities, not due to being able to access the data infrastructure but due to socioeconomic factors such as inadequate infrastructure, low national GDP, and meagre medical funding per capita .