From Collection to Analysis: A Comparison of GISAID and the Covid-19 Data Portal

Nathanael Sheehan, Sabina Leonelli, Federico Botta
University of Exeter
ns651@exeter.ac.uk

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 the Covid-19 Data Portal 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 January 2020 and January 2023, 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.

Keywords: Covid-19, genomic data sharing, data infrastructures, data governance, representativeness, Open Science

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 [1]for 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 January 2020 and January 2023, 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).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 visualisation of SARS-CoV-2 genetic variation –, 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 (Gozashti 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 science data - launched another platform for sharing scientific 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 genetic, epidemiological, socio-economic and literature 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-D place strong emphasis on the importance of sound data management and access, but their respective approaches reflect the nuanced ways in which these commitments can be interpreted and applied in practice. For CV19-DP, sound data management is predicated on the FAIR data principles, which espouse the findability, accessibility, interoperability, and reusability of research data (Wilkinson et al., 2016). This was justified in the CV19-DP publication where they state

> "...unrestricted access to data plays a critical role in the rapid coronavirus research necessary to respond to this global health crisis. This is crucial for the identification of drug targets, developing vaccines, understanding infection and symptoms, tracking the effects of new variants, and for policy makers designing public health responses. Ensuring open science and unrestricted international collaborations is of key importance, and it is recognised that these datasets must be shared openly and meet FAIR standards*"* (Harrison et al 2021)

The principles of FAIR have been enthusiastically embraced by the a number of different research areas, this is demonstrated by the growing body of literature on FAIR data sharing (Stall et al. 2019; Wise et al.2019; Bezuidenhout 2020;Goble et al. 2021;Leonelli 2021;European Commission 2022) and the cross fertilization to principles in other scientific practices such as software (Lamprecht et al. 2020; Hasselbring et al. 2020; Katz et al. 2021; Hong et al. 2022;Barker et al. 2022). 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; Tacconelli et al. 2022) and a prioritisation of machine readability over human inclusivity (Sterner and Elliot 2022).

GISAID's approach to sound data management is best understood through its database access agreement.. The EpiFlu™ Database Access Agreement is a mechanism designed to facilitate the sharing of influenza gene sequence data among researchers and public health professionals worldwide. This agreement outlines the terms under which users may provide data to the database, as well as the rights and obligations of authorised users with respect to that data. In particular, the agreement grants GISAID and authorised users a non-exclusive, worldwide, royalty-free, and irrevocable license to use, modify, display, and distribute the data submitted by users for research and intervention development purposes, provided they acknowledge the originating and submitting laboratories as the source of the data. Moreover, the agreement establishes certain restrictions on data access and distribution to ensure that users are acting in the best interests of public health. For example, users are not permitted to access or use the database in connection with any other database related to influenza gene sequences, nor are they allowed to distribute data to any third party other than authorized users. Users are also required to make best efforts to collaborate with representatives of the originating laboratory responsible for obtaining the specimen(s) and involve them in such analyses and further research using such data (GISAID, 2023). Although this agreement is established to promote collaboration between scientists, a recent publication in Science exposed GISAID as having different tiers of access which aren't defined in the agreement (Enserink and Cohen 2023).

The differences in the interpretation and implementation of sound data management by GISAID and CV19-DP illustrate the pluralistic nature of data governance and highlight the need for critical reflection on the normative foundations and ethical implications of data sharing practice. 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 (Elliot 2022). 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.

From collection to analysis: why focus on architecture, submissions and usage?-

Building on prior work on data sharing strategies and understandings of openness (Leonelli 2021; Leonelli 2023), our study takes a broad perspective on the structure and use of data infrastructures, focusing on three main characteristics: (1) *architecture* - the underlying computational systems and infrastructure that enable the functioning of data infrastructures, providing insight into the structural and organizational conditions that shape the production and dissemination of data; (2); *submissions* - the demography of contributors that are interacting with the architecture; and (3) *usage* - the lived experience of data communities who interact with data infrastructures in order to produce new outputs of knowledge. As a collective, these three components function as intermediaries in facilitating the exchange and production of SARS-CoV-2 data, thereby encapsulating the overarching role of data infrastructures as computational architectures that enable data submission and usage.

Our aim in this paper is to provide an empirical examination of a crucial process of the data journey (Leonelli and Tempini 2020), specifically the transfer of SARS-CoV-2 genetic data from the collection stage to the analytical stage. These data movements often cross institutional and international borders, thereby posing challenges to conventional scientific divisions of labor, disciplinary boundaries, and epistemic hierarchies. Despite the inherent challenges in identifying and reconstructing these journeys, they present valuable units of analysis for mapping and comparing the diverse practices and circumstances involved in the mobilization and utilization of data (Leonelli 2016A; Leonelli and Williamson 2023). At the heart of our inquiry lies the research question of how data infrastructures function as entities that mediate the interplay between data and research practices, thereby affecting the processes and outcomes of data exchange. To answer this, our methodology entails a synthesis of quantitative methods such as data collection, frequentist statistics and network analysis, and is informed by critical data studies debates on the governance and inclusivity of data infrastructures (Beaulieu et al. 2013; Kitchin 2014; Kitchin & McArdle 2016; Leonelli 2016B; Borgman 2017; Fecher 2018; Borgman and Bourne 2022; Wilson et al. 2022;Curry 2022;).

Methodology

In the forthcoming section, we explicate our methodological approach aimed at comprehensively examining the interrelated dimensions of architecture, submissions, and usage within data infrastructures, with a particular focus on the extent to which they promote openness, diversity, and representativeness. We delineate our methodology into these three key components and proceed to present a detailed account of our findings.

Architecture

In order to investigate the underlying architecture of each platform, including the data types hosted, computational systems employed, and server locations, we deploy two methods: a systematic collection of data and metadata from each platform and TCP fingerprinting.

Data and Metadata collection and mapping

On April 1st, 2023, a systematic manual collection of data and metadata was conducted across GISAID and CV-19DP in order to access and analyse these records. Data collection was conducted in accordance with the data sharing policies of both repositories (see GISAID recognition in supplementary table 1). To obtain genetic data from GISAID, we first registered and obtained access to the platform as per their terms of use. We then searched the GISAID database to access a dataset including sample collection date, location, and institutional location of sequence submissions. For the data collected from CV19-DP, we accessed the platform's top-level API to obtain all the data and metadata available.

TCP Fingerprinting

TCP/IP stack fingerprinting (OS fingerprinting) is a function often used by hackers - ethical and not so ethical - to find out the characteristics of a system they may or may not have access too. OS fingerprinting works by remotely accessing several features in the TCP/IP stack implementation and comparing them to previously defined combinations of parameters to infer matches. For the purpose of this study, we deploy this function to uncover the various degrees of openness each database is designed with, as well as the various services and steps in place to access SARS-CoV-2 metadata. The open source `whatweb` (Horton and Coles 2021) command line tool was employed to retrieve data regarding geographical location, author, type of server, and different types of plugins/libraries present in the system. The following command was used for each portal on April 1st 2023: ```./whatweb -v https://www.dataportal-url.org -a 1```, where ``-v’’ provides a verbose output of the results and ``--a 1’’ represents a soft level of pen test. The complete output for each data portal can be found in the supplementary materials.

Submission

To facilitate an empirically grounded hermeneutic understanding of the variations in data submissions across different locations, institutions, and economies, we employ a trio of data science methods to collate and visualise submissions from each platform.

Global aggregations of Sequence and Epidemiological Data

Our analysis collects over 19 million metadata points for the respective databases between the epidemiological weeks of 23 January 2020 and January 2023. Epidemiological data from the COVID-19 Data Repository by the Center for Systems Science and Engineering at Johns Hopkins University (JHU) (https://github.com/CSSEGISandData/COVID-19) is used to report on global new case counts. The data sources are linked together using country codes defined by ISO 3166-1 alpha-2. Percentages of Covid-19 cases sequenced per country - cumulatively and weekly - are calculated by filtering the data into weekly submissions and aggregating counts per country, continent and income group and region data from the World Bank (World Bank, 2023). Researchers have established a minimum sequencing requirement of 5% of SARS-CoV-2 positive cases to detect viral lineages with prevalence levels ranging from 0.1% to 1.0% (Vavrek et al., 2021). To assess the performance of GISAID and Covid-19 Data portals in meeting this requirement, we conduct an evaluation of their submissions.

Correlation between the two databases

In order to investigate the potential relationship between the number of COVID-19 data submissions to the GISAID database and the COVID-19 Data Portal (CV19-DP), a Kendall correlation test is performed. The Kendall correlation test is a robust statistical method that can provide insights into the potential relationship between two datasets, even if the relationship is not linear or if the data is non-normally distributed (Abdi 2007). This non-parametric test evaluates the association between two variables and is appropriate for non-normally distributed data. We use the number of total submissions to GISAID and CV19-DP as our variables of interest by continent. We calculate the Kendall correlation coefficient (tau) and its associated p-value using a two-tailed test with a significance level of 0.05 to evaluate the strength and significance of the correlation between the two datasets.

Geographical Data Provenance

Data provenance (often referenced as data-lineage) has been defined as the: "record trail that accounts for the origin of a piece of data (in a database, document or repository) together with an explanation of how and why it got to the present place" (Gupta 2009). Geographical data provenance is thereby concerned with tracking the locations of contributors to the inputs and outputs of computational processes (software, databases or scripts) that manipulate data across space and time for a variety of purposes. We produce a tree map visualisation for each database based on the percentage of submissions for each geographical region and country based on the total submissions as of January 2023.

Reuse

To comprehend how the wider science community employs data from either platform, we define "reuse" as the citation of the data source in publications or articles and employ various network science techniques to examine the inclusivity and diversity of publication communities (Newman 2018).

Publication networks

To glean insights into the intersection of GISAID and CV19-DP, we conduct a systematic survey of publications using the Dimensions Analytics API. Our search queries were aided by specific terms such as “coronavirus,” “SARS-CoV-2,” and “COVID-19,” in conjunction with the full name of each infrastructure. This approach ensured a targeted focus on COVID-19 research. We narrowed our results to include only articles published between January 2020 and January 2023, and then grouped them by access type, country collaboration (either single country publication SCP or multi-country publication MCP) and country of origin. We conducted summary statistics using the bibliometrix package (Aria and Cuccurullo 2017) and visualised the total publications per week, country collaboration distribution, access type distribution and mean citations per week. Next, we subjected the data to bibliographic coupling, utilizing the general formulation put forth by Kesler in 1963:

$$B=A\times A_T$$

Where A is the bipartite network adjacency matrix. The bipartite network is constructed with nodes as publication data, countries, cities, institutions and funding countries and edges as the number of couplings between two articles. Visualizations of the flow of collaborations were plotted on world maps, with dots representing nodes and edges representing the top 98% percentile of flow within the network.

Results

Architecture

TCP fingerprinting & Data and Metadata mapping

The results from the TCP fingerprinting and data/metadata mapping reveals significant differences in the architecture of each data infrastructure (see table 2). As a data consumer, one's initial point of access to the GISAID platform will likely be through its hosting on free/open-source software in Germany. In order to utilise the EpiCov3 database, which serves as the primary means of accessing GISAID data, it is necessary to register. This database is hosted on closed software in the United States and presents aggregated data in nine columns that correspond to three sub metadata categories. It is noteworthy that there is a direct mapping between the data and metadata, with each column in the aggregated data typically corresponding to a sub metadata category containing more detailed, potentially sensitive information. In addition, there is a separate metadata category that includes highly personal information about the sequence submitter.

The Covid-19 Data portal, hosted on a secure LAMP stack in the United Kingdom, serves as an alternative means of accessing sequence data. While queries made through this portal are directed to the European Nucleotide Archive database and returned as aggregated data, the portal does not directly host any data. To view the metadata associated with a particular sequence, a data user will be redirected to the ENA's application. It is worth noting that there is a disconnect between the naming conventions used for aggregated data and metadata within this system. While certain labels remain consistent, such as the accession ID, others are altered upon being pooled into the Covid-19 Data portal, e.g. "center name" (data) becomes "collection institution" (metadata). This linking strategy obscures the direct correlations between the data and the five sub metadata categories, although two of the metadata categories hosted by the ENA do contain information about linked studies, samples, and taxa. One feature found in the ENA archive, is a data provenance link to publications and samples containing the sequence.

Data Infrastructure	Country of Server	Open or Closed Software	Redirects
GISAID	Germany	Open	Y
	USA	Closed	N
CV19-DP	UK	Open	Y
	UK	Open	N

Table 1: System infrastructure (Architecture). This table details the openness and geographical location of the data infrastructure as correlated by TCP.

Data Infrastructure	**Data**	**Metadata**
GISAID	Virus name	**Institute Information**
	Accession ID	Originating lab
	Passage details/history	Address
	Collection date	Sample ID given by the originating lab
	Host	Submitting lab
	Location	Address
	Length	Sample ID given by submitting lab
	Originating lab	Authors
	Submitting lab	**Submitter Information**
		Submitter
		Submission Date
		Address
		Virus Detail
		Virus name
		Accession ID
		Type
		Clade
		Pango Lineage
		AA Substitutions
		Variant
		Passage details/history
		Sample Information
		Collection Date
		Location
		Gender
		Patient Age
		Patient Status
		Specimen source
		Sampling strategy
		Outbreak
		Last Vaccinated
		Treatment
		Sequencing Technology
		Assembly method
		Coverage
		Comment

CV19-DP	Accession Mol type Country Collection date Lineage Cross-references Release date Last modification date Center name Host Taxonomy Strain Isolate Location Coverage(%)	**Sample Attributes** ENA-CHECKLIST Center alias Center name SRA accession Submitter id Broker name Collection institution Collection date Collector name Geographic location (country and/or sea) Geographic location (region and locality) host common name Host age Host health state **Sequence** Accession Organism Mol Type Topology Base Count Dataclass Tax Division Country Collection Date Chromosome **Navigation and Cross Reference** Sample Study Taxon **Publications** Author Title **Sequence Version Archive** Accession Sequence Version First Public Last Updated Fasta | EMBL

Table 2: Data and Metadata (Architecture). This table maps outs the data to metadata relationship of each infrastructure and provides an overview of features dedicated to data provenance.

Submissions

Epidemiological surveillance of COVID-19 through sequencing efforts

The epidemiological landscape for each data infrastructure submissions vary drastically across time and space (see figure 2). The GISAID database received submissions from 197 unique countries, with 47% of these countries submitting over 5% of new cases per week. The top five countries meeting the 5% minimum were high-income countries, namely Japan, Hong Kong, Australia, United Kingdom and Canada, which had submission rates of 100%, 98.7%, 98.1%, 98.1%, and 96.8%, respectively. Regionally, the most productive regions were Europe, and North America, while other regions represented less than 15% of submissions. In terms of income group distribution, 92% of submissions to the GISAID database came from high-income countries, 5.3% from upper middle-income countries, 2.4% from lower middle-income countries, and 0.3% from low-income countries. In contrast, the Covid-19 Data portal received submissions from 117 unique countries, with only 21% of these countries representing a 5% capture of new cases. The leading countries in terms of submissions were Hong Kong, Taiwan, Lichenstein, United Kingdom and New Zealand, with submission rates of 100%, 100%, 79%, 78%, 71%, respectively. The most productive regions were from North America and Europe, while other regions represented less than 1% of submissions. The Covid-19 Data portal was heavily skewed towards high-income countries, with 98% of submissions coming from this group, while all other income groups represented less than 1% of submissions.

Geographical Data Provenance

The tree map (see figure 3) on the right illustrates that the United States, the United Kingdom, Germany, Denmark and Switzerland have been particularly active in the ENA database, comprising the majority of the total submissions with percentages of 46%, 34%, and 6.8%, 6.1%, 2.3% respectively. The remaining countries listed, including France, Australia, Iceland, and Slovakia, had much lower percentages of submissions under one percent, with all other countries representing less than 1% of the total submissions. The tree map on the left, which displays the submissions to the GISAID database, showcases a similar pattern of engagement in scientific research. The United States also leads in the number of submissions, accounting for 30.1% of the total, followed by the United Kingdom with 18.7% ,Germany with 6.7%. Japan and Denmark with 5% , France, Canada, Austria, Australia, Sweden, Spain, Brazil, Israel, Belgium, Netherlands, South Korea and Italy are also involved, with lower percentages ranging from one to two percent. The rest of the countries, including India and China, accounted for less than 0.5% of the total submissions.

Correlation between the two databases

At a global scale, we observed a strong positive correlation ($R = 0.73$, $p = < 0.001$) between the average weekly number of submissions in both databases (see figure 4). At sub-regional geographical levels (continents as determined using the United Nations dataset), we observed highly significant correlations in all regions with R scores being best fitted for Europe and North America: Africa ($R = 0.55$, $p = < 0.01$), Asia ($R = 0.69$, $p < 0.001$), Europe ($R = 0.92$), North America ($R = 0.97$, $p = < 0.01$), Oceania ($R = 0.4$, $p = < 0.1$), South America ($R = 0.31$, $p = < 0.1$) .

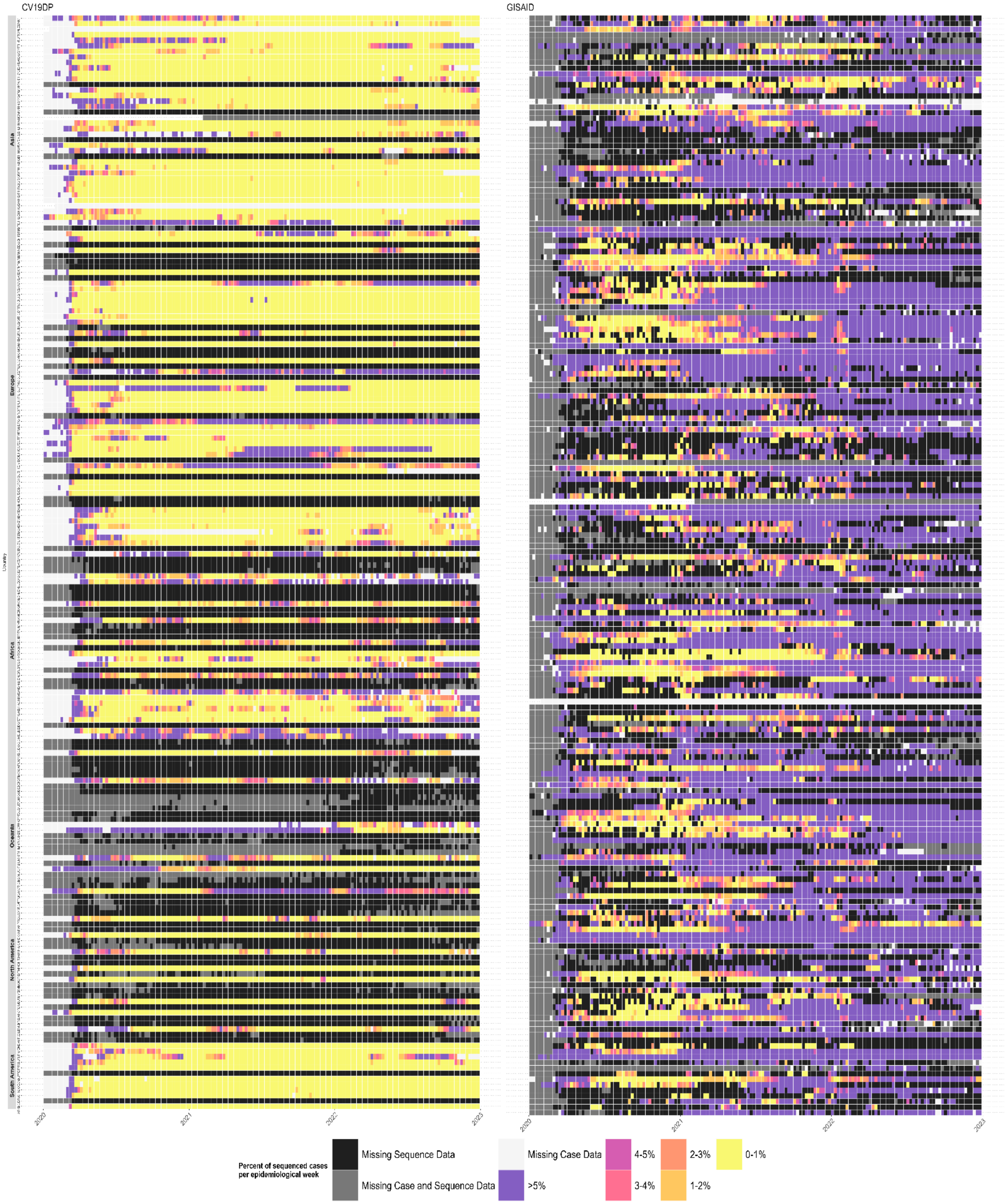

Figure 1: Epidemiological surveillance of COVID-19 through contributor efforts. On the left (CV19-DP) and the right (GISAID) displays the percentage of cases sequenced for each country and continent, organised by epidemiological week from January 2020 to January 2023.

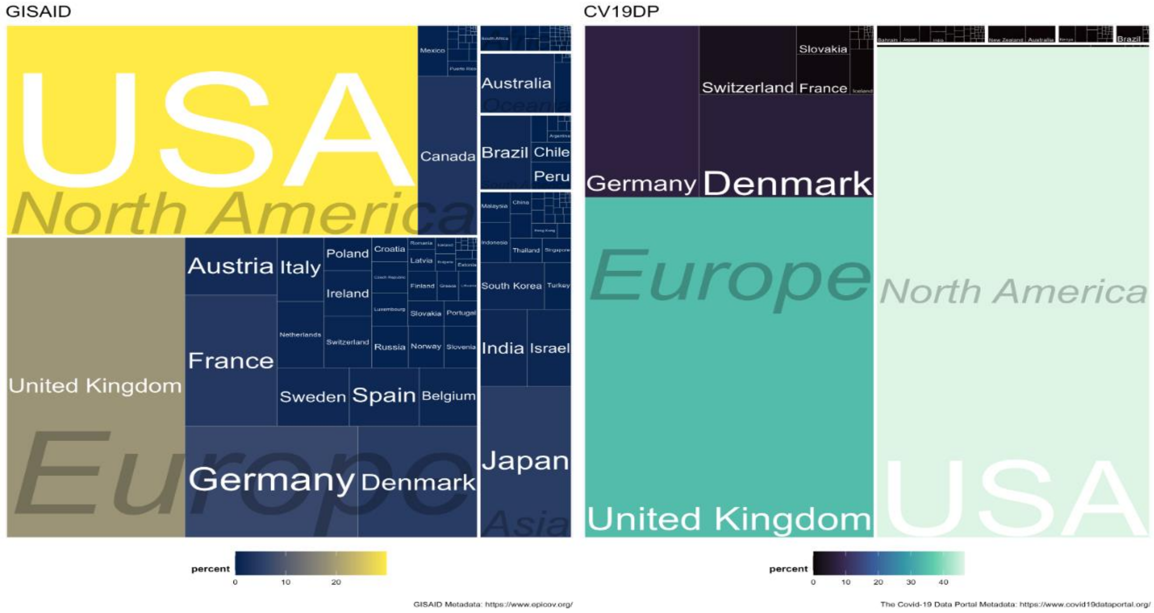

Figure 2 Database provenance (Submissions). The geographical data provenance for GISAID (left) and CV19-DP (right) is visualised as rectangles representing the percent share of the entire collection as of January 2023.

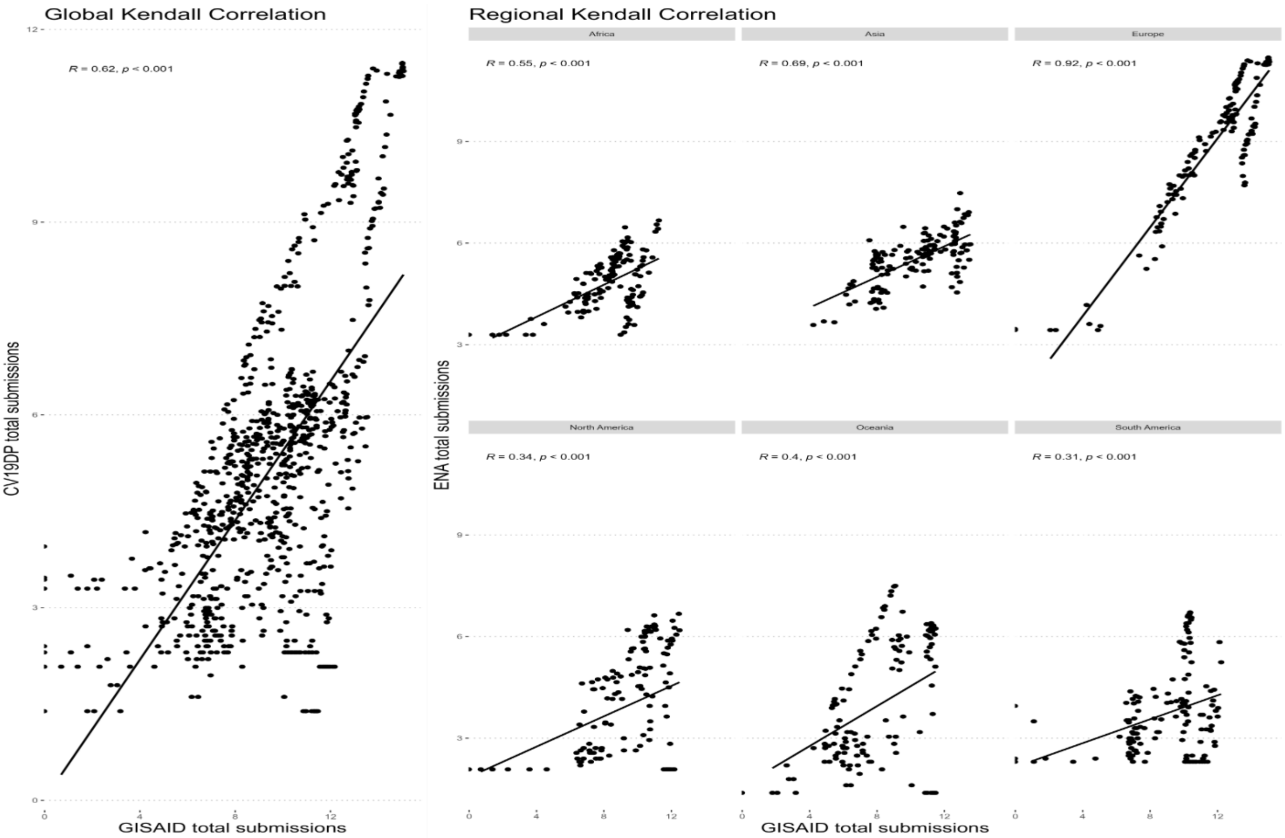

Figure 3 Database correlation (Submissions). Kendall correlation between the number of COVID-19 genome sequences submitted to GISAID and the number of COVID-19 genome sequences uploaded to the COVID-19 Data Portal (CV19-DP) as of January 2023.

Usage

Publication Networks

Following data filtering, the total number of publications mentioning GISAID was N = 9594, while publications mentioning CV19DP amounted to N = 1384.With N = 462 articles mentioning both infrastructures. Over the course of the observed timeframe, GISAID averaged a significantly higher number of articles published per week (173) compared to CV19DP (23). For both databases, the majority of publications belonged to the gold or green access category, followed by hybrid and bronze access types, with closed access publications forming the smallest group. Analysis of the countries contributing to the dataset revealed that the United States had the highest number of articles (2175), followed by China (1140), India (646), United Kingdom (632), and Italy (503). Among 22 countries analysed, 19 had a higher percentage of articles in GISAID than in ENA, except for Iran, Turkey, and Egypt, for both MCP and SCP. Multi-country partnerships were the most common type of collaboration, accounting for 78.3% of all partnerships, while single country partnerships accounted for the remaining 22.7%. The mean citations per week for each article demonstrated that GISAID had a slightly higher average citation rate (40.2) compared to CV19DP (36.5). Nevertheless, it is noteworthy that CV19DP had more citations on many occasions throughout the observation period.

The results from the network analysis displayed both homogenous and heterogeneous characteristics between the number of nodes in a given country, the leading 98% share of flow being circulated between the network (see figure 6) and network metrics (table 4 and 5). Within the funding countries network (figure 6 - first row), the leading flow between countries is only present (> 153) between the United States and the United Kingdom for CV19DP, however for GISAID the funding network spreads from North and South America to Northern Europe and East Asia (> 303.4). Within this network, the GISAID network featuring eight extra nodes – in South America, Asia and Europe~~)~~ - and slightly higher network scores for all other measurements. In the countries network, both GISAID and CV19DP networks exhibited relatively high-density values, with GISAID having a higher transitivity value. Both networks had low diameter and distance values, indicating that they are relatively localised with a high degree of connectivity between nodes. Within this network the leading flow for GISAID (157.5) and CV19DP (118.92) follow similar connection patterns (figure 6 – second row). The institutions network showed that the GISAID network had a significantly larger size (5193 nodes) than the ENA network (2163 nodes), although the CV19DP was found to be more tightly interconnected, as evidenced by its higher transitivity. This interconnectedness is also apparent~~ in~~ the leading flow for the network (figure 6 – third row), which shows similar migration patterns of the network but deeply different percentiles, GISAID (7) and CV19DP (2). In the cities network, GISAID had a larger size with 1930 nodes compared to ENA's 1101 nodes. Notably, the GISAID network leading flow (22) covered a much wider and inclusive landscape than CV19DP (12).

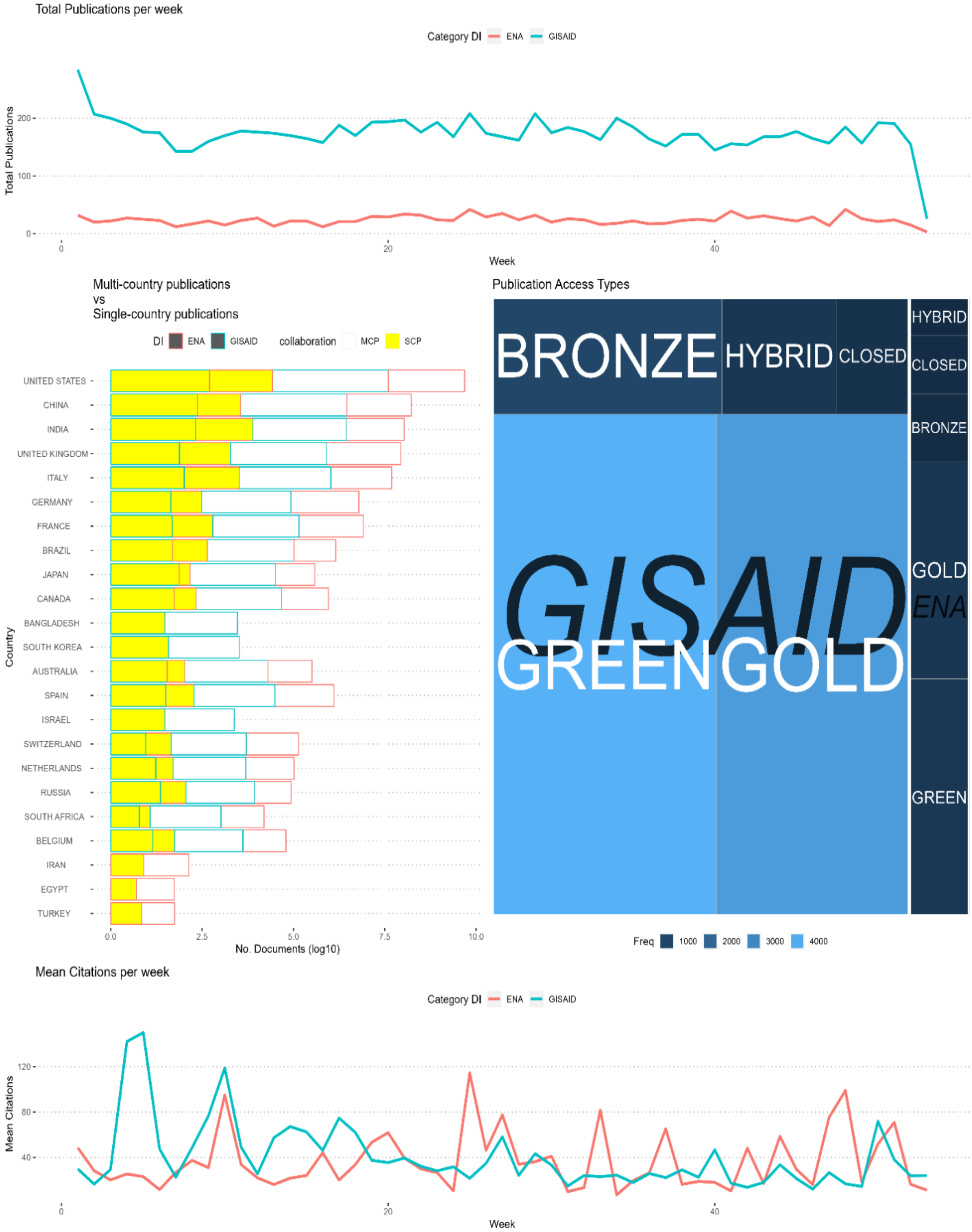

Figure 4 Summary of bibliometric analysis (Usage). Top – Total publication per week. Middle left - Single Country Publication vs Multi Country Publication. Middle-right – Frequency distribution of access types. Bottom – Mean citations of publications per week.

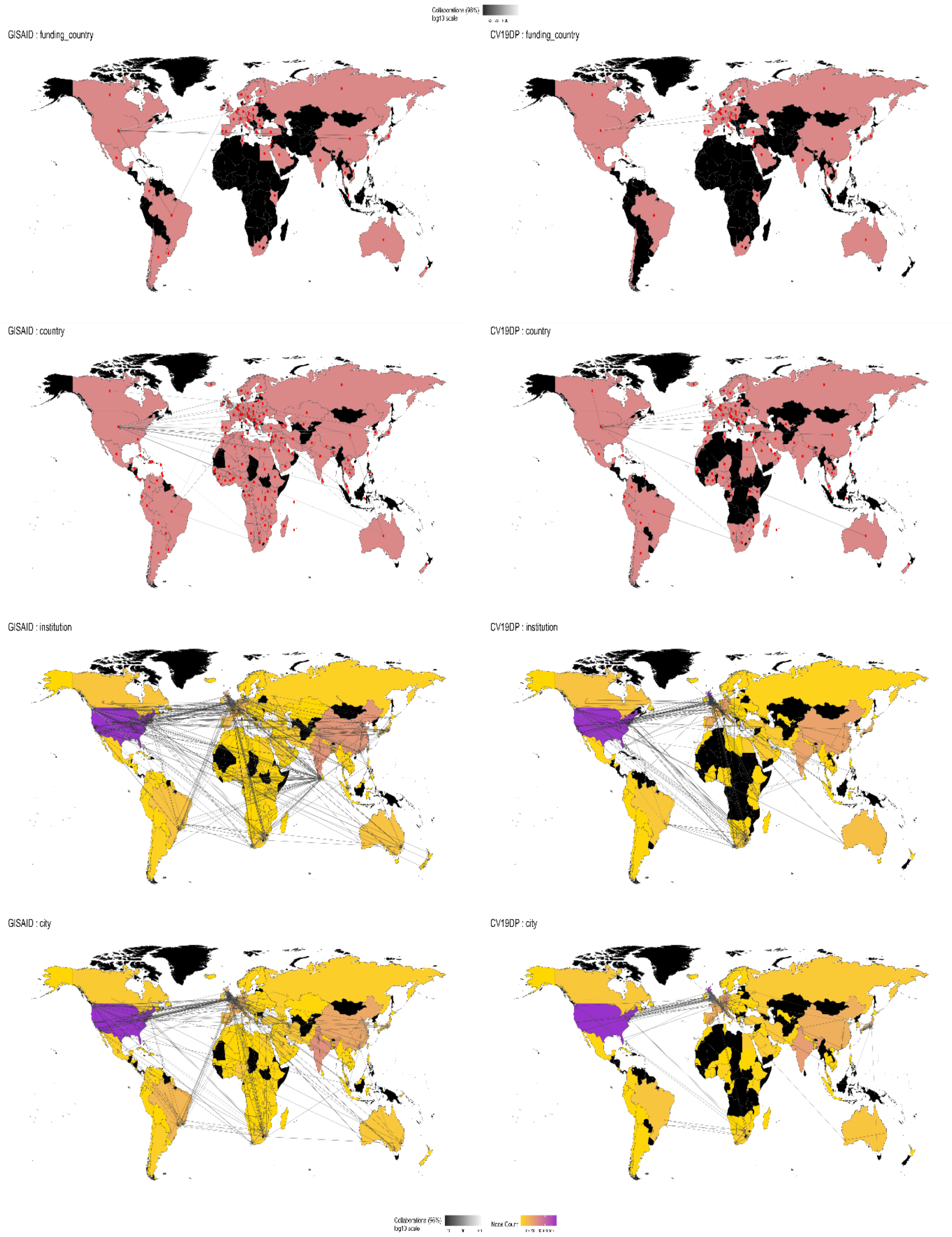

Figure 5 Collaboration Networks (Usage). These maps illustrate the top 98% flow of collaboration and node destinations for GISAID (left) and CV19DP (right). In the top two rows, nodes are represented as red dots and display the funding countries (top) and countries (bottom) involved in the collaboration. In the bottom two rows, nodes are visualised as choropleths, indicating the volume of nodes in a given country for the institutions (top) and cities (bottom). For all maps, the leading 98% flow of edges is displayed in grey, emphasizing the most significant collaborations.

Size	density	transitivity	diameter	distance	avgpath	categories
5193	0	0.36	9	0.07	3.83	"institution"
1877	0.01	0.36	6	0.14	3.74	"city"
145	0.13	0.66	5	0.29	4.47	"country"
54	0.12	0.52	4	0.34	3.9	"funding country"

Table 3: Network Statistics GISAID. Network statistics of publication data from the Global Initiative on Sharing All Influenza Data

Size	density	transitivity	diameter	distance	avgpath	categories
2163	0.01	0.9	11	0.1	3.7	"institution"
951	0.02	0.84	6	0.17	3.72	"city"
99	0.09	0.51	4	0.29	3.53	"country"
46	0.09	0.4	4	0.34	4.13	"funding country"

Table 3: Network Statistics CV91-DP. Network statistics of publication data from the Global Initiative on Sharing All Influenza Data

Discussion

Upon conducting our initial mapping of the architecture of each database, we observed that GISAID's architecture is a complex mixture of open and proprietary tools located across different geographical localities, with servers split between Germany and the United States of America. In contrast, CV19-DP is more geographically situated, operating solely in the United Kingdom and hosted exclusively on open servers. These findings challenge oversimplified dichotomies of open versus closed systems and highlight the intricate interplay between open and proprietary software in enabling open data sharing (Arita 2021; Leonelli 2021). Furthermore, our analysis of the data provided by each infrastructure has revealed further distinctions in their architectures, with GISAID adopting a more standardised approach to data management than CV19-DP, albeit with limitations in supporting certain aspects of data provenance such as versioning of a genome or linkage to publications. While both infrastructures integrate tools for users to explore sequence data, GISAID offers more extensive and frequently updated analysis tools. Notably, one significant difference between the two platforms is the types of data hosted and the authentication required to access such data. It is evident that CV19-DP houses a broader range of epidemiological data without authentication, while GISAID solely focuses on sharing RNA or protein data. Thus, both models of data sharing have their respective strengths and limitations but are in turn built for slightly different purposes: GISAID’s architecture functions as a secure entry point for collaborative research to be done in a specific domain, with key figures/analysis being the central attraction for users of the platform. CV19-DP architecture, on the other hand, serves more as a middle ground between the number of data sources in which already exist as utilities in viral data sharing and epidemiology more generally, while this means there is a less of a concentration on tools/figures it permits users of the infrastructure to explore their own research questions and methods by linking together the wider variety of research data types.

Through our analysis of submission efforts, we are confronted with highly unequal SARS-CoV-2 genomic surveillance efforts at the global scale, which highlights the structural inequalities inherent in our current scientific systems (Brito et al. 2021; Tegally et al.

2022). Although GISAID received submissions from 194 countries, with many meeting the minimum 5% surveillance requirement, the data provenance of both infrastructures is overwhelmingly skewed toward high-income countries. Specifically, GISAID and CV19-DP exhibit a 93% and 98% representation of data from high-income countries, respectively. These large percentages raise concerns about the representation of data from low- and middle-income countries within both data infrastructures, which raises important questions about access to and sharing of scientific resources, as well as the potential for biases in data sampling based on incomplete data. Efforts to improve data sharing and promote equity in scientific research are critical for ensuring that all populations, regardless of geography or economic status, have access to the best available information and resources for preventing and treating disease (Prat and Bull 2021; Staunton et al. 2021). In this sense, our findings indicate that the GISAID model is more effective in supporting submissions by incorporating constraints on data access and reuse, thereby building trustworthiness among users from different resource environments to collaborate and engage in data generation (ODI, 2019; RDA, 2020). These results reinforce claims by Bernasconi (2021) that a partially closed access model is preferred at a global scale for viral data sharing.

In our study of data reuse, we have uncovered a multifaceted interplay of factors that influence the production and dissemination of scientific information. The observed discrepancies in the total number of publications referencing GISAID and CV19DP after data filtering raises important questions about the quality of these databases. One might ask, for instance, whether GISAID's significantly higher number of articles published per week reflects a more robust and comprehensive approach to data collection and analysis, or whether it is simply indicative of a less rigorous and more inclusive process of data curation? Similarly, the prevalence of gold and green access categories in both databases, along with the relatively small proportion of closed access publications, suggests a broader trend towards greater transparency and openness in data science research, while also pointing to potential concerns around accessibility and equitable distribution of research resources. The analysis of countries contributing to the dataset provides a valuable insight into the global distribution of COVID-19 research, with the United States, China, India, United Kingdom, and Italy emerging as the top contributors. The dominance of multi-country partnerships in research collaborations, however, raises questions about the extent to which research data accumulated in centralised data infrastructures can truly be considered a global endeavour, or whether it is still largely driven by a select few countries and institutions. The results of the network analysis offer further insights into the patterns of collaboration and flow of information withinSARS-CoV-2 research. The heterogeneity observed between the funding countries network for GISAID and CV19DP, for instance, raises important questions about the extent to which funding sources influence the distribution and accessibility of research resources. Moreover, the relatively high-transitivity values and low diameter and distance values observed in both the countries and institutions network for CV19DP suggest a highly interconnected and localised research landscape. This localised nature of research collaborations in data science could either reflect an important trend towards the formation of smaller and more focused research communities, or potentially indicate a limited and self-referential approach to data reuse and analysis.

	Architecture	Submissions	Usage
GISAID	**PROS:** Authentication system for users Clear mapping between data and metadata Daily updated visualisations Analysis tools to explore sequence data.	**PROS:** Representation of data from low- and middle-income countries Large share of countries achieving the recommended 5% SARS-CoV-2 surveillance baseline	**PROS:** Large share of multi country publications Large volume of articles published. Fewer articles published in closed access journals. Majority number of articles published as open access. Larger network of authors, cities, institutions, countries and funders.
	CONS: Mix of open and closed software. Distributed across geographies. Limited tooling for data provenance.		**CONS:** Smaller transitivity in network

Table 4. Summary of findings for GISAID

	Architecture	**Submissions**	**Usage**
CV19-DP	**PROS:** Servers all hosted on FOSS software. Additional features of data provenance Analysis tools to explore sequence data. Other forms of data relevant to the study of epidemiology Situated in one geography.	**PROS:** Partial share of countries achieving the recommended 5% SARS-CoV-2 surveillance baseline.	**PROS:** Larger share of single country publications Majority number of articles published as open access. Greater transitivity in network.
	CONS: No authentication system Less clear mapping between data and metadata	**CONS:** Representation of data from mostly high-income countries Majority share of countries not achieving the recommended 5% SARS-CoV-2 surveillance baseline	**CONS:** Smaller network of authors, cities, institutions, countries and funders Smaller volume of articles published. Greater number of articles published in closed access journals.

Table 5. Summary of findings for CV19-DP

Conclusion

Our analysis highlights the strengths and limitations of two distinct approaches to open data governance embodied by GISAID and CV19-DP. These databases furnish valuable resources for scientific research, yet they diverge in their architecture, submissions and reuse of data from different countries and labs. For CV19-DP the most pressing issue is to enhance trust and thereby representativeness of submissions, while in turn accelerating the pace of research. This may come in the form of requiring registration to better track data usage or investment in governance of prospective uses. All of which improves representativeness and participation in data interpretation, which in turn improves actionability. While for GISAID, the most feasible improvement would be in their architectures support of data provenance, such as noting changes to the original sequence, as well as linking the sequence to publications or samples. Rather than framing these two approaches as antithetical to one and other, we rather aim to conclude that the virtues from either infrastructure can be brought together to establish a more effective framework of open data governance more broadly. By cultivating open data governance and fostering collaboration, we can enhance our collective comprehension of the ongoing crises and develop more effective interventions and treatments for the betterment of all.

Appendix

Supplementary files 1. List of GISAID tools and figures

Tools and description	Figures
Audacity – Global phylogeny of hCoV-19 as a downloadable newick tree file. AudacityInstant - Query using a sample seqeunce Epi-CoV database for similarity **BLAST** – Query using a sample sequence against the Epi-CoV database collected from the last three months **CoVizue** – Near real-time visulsation of SARS-CoV-2 genetic variation **Emerging Variants** – A dashboard with maps and visualisations for each lineage of SARS-CoV-2. **Lineage Frequency** – A dashboard visualising lineage distribution Official Reference Sequence – High quality FASTA file of the representative of the early outbreak sequences **PrimerChecker** – Search tool for short sequence match to search primers in the Epi-CoV database **Submission tracker** – A world map depicting submission rates to the Epi-CoV database **Spike glycoprotein mutation surveillance** - Spike glycoprotein variation altering potential N-glycosylation sites lineage emerging mutations **Wasterwater** – A subset of the Epi-Cov database coming from wastewater surveillance	• Representative phylogenetics of recent genome sequences • Representative phylogenetics of recent genome sequences • Global Occurrence and Frequencies of VOIs • Global Occurrence and Frequencies of VOIs • Global Occurrence and Frequencies of VUMs • Global Occurrence and Frequencies of VUMs • Phylodynamic of VOIs and VUMs • Phylodynamic of VOIs and VUMs • Emerging variants by Spread • Emerging variants by Acceleration • Full genome tree derived from all outbreak sequences • Time course of variant distribution in all submitted sequences • Regional distribution of variants in collected sequences • Time course of Omicron variant sub lineage distribution • Regional trends of Omicron variant sub lineages in collected sequences • Regional distribution of variants in new sequences • Time course of clade distribution in collected sequences • Regional clade distribution of new sequences • Distribution of collection dates of new sequences • Change in proportions of Omicron sub lineages • Breakdown of new sequences by clade, then by territory • Spike glycoprotein mutation surveillance • Receptor binding surveillance for complete genomes • Common primer check for high quality genomes

Reproducibility

The code and data that support the findings of this study are openly available on the following links:

- https://github.com/natesheehan/OPEN-GM/code/
- https://github.com/natesheehan/OPEN-GM/data/

The findings of this study were analysed using the R programming language and RStudio. Within the README file of the repository are instructions on how to fully reproduce the results, once again from collection to analysis.

Acknowledgements

We would like to thank the authors from the originating labs - together with the submitting labs - who were responsible for collecting, codifying and sharing the pathogen related data to which this analysis relies on. To this end, an acknowledgement table can be found from gisaid.org with the following set accession id: EPI_SET_220927xo at this URL https://github.com/natesheehan/OPEN-GM/blob/main/gisaid_supplemental_table_epi_set_220927xo.pdf

We would also like to thank the Covid-19 Data Portal and Our World in Data team for their efforts in collecting and curating the epidemiological data used in this study.

Funding Information

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.

Competing Interests

The authors have no competing interests to declare.

References

Abdi, H., n.d. The Kendall Rank Correlation Coefficient.

Aria, M., Cuccurullo, C., 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics 11, 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Arita, M., 2021. Open Access and Data Sharing of Nucleotide Sequence Data. Data Science Journal 20, 28. https://doi.org/10.5334/dsj-2021-028

Barker, M., Chue Hong, N.P., Katz, D.S., Lamprecht, A.-L., Martinez-Ortiz, C., Psomopoulos, F., Harrow, J., Castro, L.J., Gruenpeter, M., Martinez, P.A., Honeyman, T., 2022. Introducing the FAIR Principles for research software. Sci Data 9, 622. https://doi.org/10.1038/s41597-022-01710-x

Beaulieu, A., Scharnhorst, A., Wouters, P., Wyatt, S. (Eds.), 2013. Virtual knowledge: experimenting in the humanities and the social sciences. The MIT Press, Cambridge, Massachusetts.

Bernasconi, A., Canakoglu, A., Masseroli, M., Pinoli, P., Ceri, S., 2021. A review on viral data sources and search systems for perspective mitigation of COVID-19. Briefings in Bioinformatics 22, 664–675. https://doi.org/10.1093/bib/bbaa359

Bezuidenhout, L., 2020. Being Fair about the Design of FAIR Data Standards. Digit. Gov.: Res. Pract. 1, 18:1-18:7. https://doi.org/10.1145/3399632

Bezuidenhout, L., Chakauya, E., 2018. Hidden concerns of sharing research data by low/middle-income country scientists. Glob Bioeth 29, 39–54. https://doi.org/10.1080/11287462.2018.1441780

Bibliographic coupling between scientific papers - Kessler - 1963 - American Documentation - Wiley Online Library [WWW Document], n.d. URL https://onlinelibrary.wiley.com/doi/10.1002/asi.5090140103 (accessed 5.9.23).

Big Data, new epistemologies and paradigm shifts - Rob Kitchin, 2014 [WWW Document], n.d. URL https://journals.sagepub.com/doi/10.1177/2053951714528481 (accessed 3.2.23).

Boeckhout, M., Zielhuis, G.A., Bredenoord, A.L., 2018. The FAIR guiding principles for data stewardship: fair enough? Eur J Hum Genet 26, 931–936. https://doi.org/10.1038/s41431-018-0160-0

Borgman, C.L., 2017. Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press.

Brito, A.F., Semenova, E., Dudas, G., Hassler, G.W., Kalinich, C.C., Kraemer, M.U.G., Ho, J., Tegally, H., Githinji, G., Agoti, C.N., Matkin, L.E., Whittaker, C., Bulgarian SARS-CoV-2 sequencing group, Communicable Diseases Genomics Network (Australia and New Zealand), COVID-19 Impact Project, Danish Covid-19 Genome Consortium, Fiocruz COVID-19 Genomic Surveillance Network, GISAID core curation team, Network for Genomic Surveillance in South Africa (NGS-SA), Swiss SARS-CoV-2 Sequencing Consortium, Howden, B.P., Sintchenko, V., Zuckerman, N.S., Mor, O., Blankenship, H.M., de Oliveira, T., Lin, R.T.P., Siqueira, M.M., Resende, P.C., Vasconcelos, A.T.R., Spilki, F.R., Aguiar, R.S., Alexiev, I., Ivanov, I.N., Philipova, I., Carrington, C.V.F., Sahadeo, N.S.D., Branda, B., Gurry, C., Maurer-Stroh, S., Naidoo, D., von Eije, K.J., Perkins, M.D., van Kerkhove, M., Hill, S.C., Sabino, E.C., Pybus, O.G., Dye, C., Bhatt, S., Flaxman, S., Suchard, M.A., Grubaugh, N.D., Baele, G., Faria, N.R., 2022. Global disparities in SARS-CoV-2 genomic surveillance. Nat Commun 13, 7003. https://doi.org/10.1038/s41467-022-33713-y

Butler, D., 2009. Flu database rocked by legal row. Nature 460, 787–787. https://doi.org/10.1038/460786b

Chen, Z., Azman, A.S., Chen, X., Zou, J., Tian, Y., Sun, R., Xu, X., Wu, Y., Lu, W., Ge, S., Zhao, Z., Yang, J., Leung, D.T., Domman, D.B., Yu, H., 2022. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. Nat Genet 54, 499–507. https://doi.org/10.1038/s41588-022-01033-y

Curry, H.A., 2022. The history of seed banking and the hazards of backup. Soc Stud Sci 52, 664–688. https://doi.org/10.1177/03063127221106728

Directorate-General for Research and Innovation (European Commission), Maxwell, L., 2022. Maximising investments in health research: FAIR data for a coordinated COVID 19 response : workshop report. Publications Office of the European Union, LU.

Elbe, S., Buckland-Merrett, G., 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall 1, 33–46. https://doi.org/10.1002/gch2.1018

Elliott, K.C., 2022. Values in Science, Elements in the Philosophy of Science. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781009052597

Fecher, B., 2018. Eine Reputationsökonomie. Springer Fachmedien, Wiesbaden. https://doi.org/10.1007/978-3-658-20895-0

Goble, C., Soiland-Reyes, S., Bacall, F., Owen, S., Williams, A., Eguinoa, I., Droesbeke, B., Leo, S., Pireddu, L., Rodríguez-Navas, L., 2021. Implementing FAIR digital objects in the EOSC-Life workflow collaboratory. Zenodo.

Gozashti, L., Corbett-Detig, R., 2021. Shortcomings of SARS-CoV-2 genomic metadata. BMC Res Notes 14, 189. https://doi.org/10.1186/s13104-021-05605-9

Greenemeier, L., n.d. Open-Access Flu Research Web Site Is Relaunched Amid Controversy [WWW Document]. Scientific American. URL https://www.scientificamerican.com/article/gisaid-sib-flu-database/ (accessed 2.21.23).

Gupta, A., 2009. Data Provenance, in: LIU, L., ÖZSU, M.T. (Eds.), Encyclopedia of Database Systems. Springer US, Boston, MA, pp. 608–608. https://doi.org/10.1007/978-0-387-39940-9_1305

Harrison, P.W., Lopez, R., Rahman, N., Allen, S.G., Aslam, R., Buso, N., Cummins, C., Fathy, Y., Felix, E., Glont, M., Jayathilaka, S., Kadam, S., Kumar, M., Lauer, K.B., Malhotra, G., Mosaku, A., Edbali, O., Park, Y.M., Parton, A., Pearce, M., Estrada Pena, J.F., Rossetto, J., Russell, C., Selvakumar, S., Sitjà, X.P., Sokolov, A., Thorne, R., Ventouratou, M., Walter, P., Yordanova, G., Zadissa, A., Cochrane, G., Blomberg, N., Apweiler, R., 2021. The COVID-19 Data Portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. Nucleic Acids Res 49, W619–W623. https://doi.org/10.1093/nar/gkab417

Hasselbring, W., Carr, L., Hettrick, S., Packer, H., Tiropanis, T., 2020. From FAIR research data toward FAIR and open research software. it - Information Technology 62, 39–47. https://doi.org/10.1515/itit-2019-0040

Hong, N.P.C., Katz, D.S., Barker, M., Lamprecht, A.-L., Martinez, C., Psomopoulos, F.E., Harrow, J., Castro, L.J., Gruenpeter, M., Martinez, P.A., Honeyman, T., Struck, A., Lee, A., Loewe, A., Werkhoven, B. van, Garijo, D., Plomp, E., Genova, F., Shanahan, H., Hellström, M., Sandström, M., Sinha, M., Kuzak, M., Herterich, P., Islam, S., Sansone, S.-A., Pollard, T., Atmojo, U.D., Williams, A., Czerniak, A., Niehues, A., Fouilloux, A.C., Desinghu, B., Goble, C., Richard, C., Gray, C., Erdmann, C., Nüst, D., Tartarini, D., Ranguelova, E., Anzt, H., Todorov, I., McNally, J., Burnett, J., Garrido-Sánchez, J., Belhajjame, K., Sesink, L., Hwang, L., Tovani-Palone, M.R., Wilkinson, M.D., Servillat, M., Liffers, M., Fox, M., Miljković, N., Lynch, N., Lavanchy, P.M., Gesing, S., Stevens, S., Cuesta, S.M., Peroni, S., Soiland-Reyes, S., Bakker, T., Rabemanantsoa, T., Sochat, V., Yehudi, Y., Wg, F., 2022. FAIR Principles for Research Software (FAIR4RS Principles). https://doi.org/10.15497/RDA00065

Kalia, K., Saberwal, G., Sharma, G., 2021. The lag in SARS-CoV-2 genome submissions to GISAID. Nat Biotechnol 39, 1058–1060. https://doi.org/10.1038/s41587-021-01040-0

Katz, D.S., Gruenpeter, M., Honeyman, T., 2021. Taking a fresh look at FAIR for research software. Patterns 2, 100222. https://doi.org/10.1016/j.patter.2021.100222

Khare, S., Gurry, C., Freitas, L., Schultz, M.B., Bach, G., Diallo, A., Akite, N., Ho, J., Lee, R.T., Yeo, W., Curation Team, G.C., Maurer-Stroh, S., 2021. GISAID's Role in Pandemic Response. China CDC Wkly 3, 1049–1051. https://doi.org/10.46234/ccdcw2021.255

Kitchin, R., McArdle, G., 2016. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society 3, 2053951716631130. https://doi.org/10.1177/2053951716631130

Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., Dominguez Del Angel, V., van de Sandt, S., Ison, J., Martinez, P.A., McQuilton, P., Valencia, A., Harrow, J., Psomopoulos, F., Gelpi, J.L., Chue Hong, N., Goble, C., Capella-Gutierrez, S., 2020. Towards FAIR principles for&nbsp;research&nbsp;software. Data Science 3, 37–59. https://doi.org/10.3233/DS-190026

Leonelli, S., 2023. Philosophy of Open Science [WWW Document]. URL http://philsci-archive.pitt.edu/21986/ (accessed 5.5.23).

Leonelli, S., 2016. Data-Centric Biology: A Philosophical Study. University of Chicago Press, Chicago, IL.

Leonelli, S., Lovell, R., Wheeler, B.W., Fleming, L., Williams, H., 2021. From FAIR data to fair data use: Methodological data fairness in health-related social media research. https://doi.org/10.1177/20539517211010310

Open letter: Support data sharing for COVID-19 [WWW Document], n.d. URL https://www.covid19dataportal.org/support-data-sharing-covid19 (accessed 2.21.23).

Re3data.Org, 2012a. GISAID. https://doi.org/10.17616/R3Q59F

Re3data.Org, 2012b. GISAID. https://doi.org/10.17616/R3Q59F

Saravanan, K.A., Panigrahi, M., Kumar, H., Rajawat, D., Nayak, S.S., Bhushan, B., Dutt, T., 2022. Role of genomics in combating COVID-19 pandemic. Gene 823, 146387. https://doi.org/10.1016/j.gene.2022.146387

Science Magazine - Control issues [WWW Document], n.d. URL https://www.sciencemagazinedigital.org/sciencemagazine/library/item/28_april_2023/4098017/?Cust_No=40579050&utm_source=newsletter&utm_medium=email&utm_campaign=TXSCI2230427002&utm_content=gtxcel (accessed 5.9.23).

Shu, Y., McCauley, J., 2017. GISAID: Global initiative on sharing all influenza data – from vision to reality. Euro Surveill 22, 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., Parsons, M., Robinson, E., Wyborn, L., 2019. Make scientific data FAIR. Nature 570, 27–29. https://doi.org/10.1038/d41586-019-01720-7

Sterner, B., Elliott, S., n.d. The FAIR and CARE Data Principles Influence Who Counts As a Participant in Biodiversity Science by Governing the Fitness-for-Use of Data.

Tacconelli, E., Gorska, A., Carrara, E., Davis, R.J., Bonten, M., Friedrich, A.W., Glasner, C., Goossens, H., Hasenauer, J., Abad, J.M.H., Peñalvo, J.L., Sanchez-Niubo, A., Sialm, A., Scipione, G., Soriano, G., Yazdanpanah, Y., Vorstenbosch, E., Jaenisch, T., 2022a. Challenges of data sharing in European Covid-19 projects: A learning opportunity for advancing pandemic preparedness and response. The Lancet Regional Health – Europe 21. https://doi.org/10.1016/j.lanepe.2022.100467

Tacconelli, E., Gorska, A., Carrara, E., Davis, R.J., Bonten, M., Friedrich, A.W., Glasner, C., Goossens, H., Hasenauer, J., Abad, J.M.H., Peñalvo, J.L., Sanchez-Niubo, A., Sialm, A., Scipione, G., Soriano, G., Yazdanpanah, Y., Vorstenbosch, E., Jaenisch, T., 2022b. Challenges of data sharing in European Covid-19 projects: A learning opportunity for advancing pandemic preparedness and response. The Lancet Regional Health – Europe 21. https://doi.org/10.1016/j.lanepe.2022.100467

Tegally, H., San, J.E., Cotten, M., Moir, M., Tegomoh, B., Mboowa, G., Martin, D.P., Baxter, C., Lambisia, A.W., Diallo, A., Amoako, D.G., Diagne, M.M., Sisay, A., Zekri, A.-R.N., Gueye, A.S., Sangare, A.K., Ouedraogo, A.-S., Sow, A., Musa, A.O., Sesay, A.K., Abias, A.G., Elzagheid, A.I., Lagare, A., Kemi, A.-S., Abar, A.E., Johnson, A.A., Fowotade, A., Oluwapelumi, A.O., Amuri, A.A., Juru, A., Kandeil, A., Mostafa, A., Rebai, A., Sayed, A., Kazeem, A., Balde, A., Christoffels, A., Trotter, A.J., Campbell, A., Keita, A.K., Kone, A., Bouzid, A., Souissi, A., Agweyu, A., Naguib, A., Gutierrez, A.V., Nkeshimana, A., Page, A.J., Yadouleton, A., Vinze, A., Happi, A.N., Chouikha, A., Iranzadeh, A., Maharaj, A., Batchi-Bouyou, A.L., Ismail, A., Sylverken, A.A., Goba, A., Femi, A., Sijuwola, A.E., Marycelin, B., Salako, B.L., Oderinde, B.S., Bolajoko, B., Diarra, B., Herring, B.L., Tsofa, B., Lekana-Douki, B., Mvula, B., Njanpop-Lafourcade, B.-M., Marondera, B.T., Khaireh, B.A., Kouriba, B., Adu, B., Pool, B., McInnis, B., Brook, C., Williamson, C., Nduwimana, C., Anscombe, C., Pratt, C.B., Scheepers, C., Akoua-Koffi, C.G., Agoti, C.N., Mapanguy, C.M., Loucoubar, C., Onwuamah, C.K., Ihekweazu, C., Malaka, C.N., Peyrefitte, C., Grace, C., Omoruyi, C.E., Rafaï, C.D., Morang'a, C.M., Erameh, C., Lule, D.B., Bridges, D.J., Mukadi-Bamuleka, D., Park, D., Rasmussen, D.A., Baker, D., Nokes, D.J., Ssemwanga, D., Tshiabuila, D., Amuzu, D.S.Y., Goedhals, D., Grant, D.S., Omuoyo, D.O., Maruapula, D., Wanjohi, D.W., Foster-Nyarko, E., Lusamaki, E.K., Simulundu, E., Ong'era, E.M., Ngabana, E.N., Abworo, E.O., Otieno, E., Shumba, E., Barasa, E., Ahmed, E.B., Ahmed, E.A., Lokilo, E., Mukantwari, E., Philomena, E., Belarbi, E., Simon-Loriere, E., Anoh, E.A., Manuel, E., Leendertz, F., Taweh, F.M., Wasfi, F., Abdelmoula, F., Takawira, F.T., Derrar, F., Ajogbasile, F.V., Treurnicht, F., Onikepe, F., Ntoumi, F., Muyembe, F.M., Ragomzingba, F.E.Z., Dratibi, F.A., Iyanu, F.-A., Mbunsu, G.K., Thilliez, G., Kay, G.L., Akpede, G.O., van Zyl, G.U., Awandare, G.A., Kpeli, G.S., Schubert, G., Maphalala, G.P., Ranaivoson, H.C., Omunakwe, H.E., Onywera, H., Abe, H., Karray, H., Nansumba, H., Triki, H., Kadjo, H.A.A., Elgahzaly, H., Gumbo, H., Mathieu, H., Kavunga-Membo, H., Smeti, I., Olawoye, I.B., Adetifa, I.M.O., Odia, I., Ben Boubaker, I.B., Muhammad, I.A., Ssewanyana, I., Wurie, I., Konstantinus, I.S., Halatoko, J.W.A., Ayei, J., Sonoo, J., Makangara, J.-C.C., Tamfum, J.-J.M., Heraud, J.-M., Shaffer, J.G., Giandhari, J., Musyoki, J., Nkurunziza, J., Uwanibe, J.N., Bhiman, J.N., Yasuda, J., Morais, J., Kiconco, J., Sandi, J.D., Huddleston, J., Odoom, J.K., Morobe, J.M., Gyapong, J.O., Kayiwa, J.T., Okolie, J.C., Xavier, J.S., Gyamfi, J., Wamala, J.F., Bonney, J.H.K., Nyandwi, J., Everatt, J., Nakaseegu, J., Ngoi, J.M., Namulondo, J., Oguzie, J.U., Andeko, J.C., Lutwama, J.J., Mogga, J.J.H., O'Grady, J., Siddle, K.J., Victoir, K., Adeyemi, K.T., Tumedi, K.A., Carvalho, K.S., Mohammed, K.S., Dellagi, K., Musonda, K.G., Duedu, K.O., Fki-Berrajah, L., Singh, L., Kepler, L.M., Biscornet, L., de Oliveira Martins, L., Chabuka, L., Olubayo, L., Ojok, L.D., Deng, L.L., Ochola-Oyier, L.I., Tyers, L., Mine, M., Ramuth, M., Mastouri, M., ElHefnawi, M., Mbanne, M., Matsheka, M.I., Kebabonye, M., Diop, M., Momoh, M., Lima Mendonça, M. da L., Venter, M., Paye, M.F., Faye, M., Nyaga, M.M., Mareka, M., Damaris, M.-M., Mburu, M.W., Mpina, M.G., Owusu, M., Wiley, M.R., Tatfeng, M.Y., Ayekaba, M.O., Abouelhoda, M., Beloufa, M.A., Seadawy, M.G., Khalifa, M.K., Matobo, M.M., Kane, M., Salou, M., Mbulawa, M.B., Mwenda, M., Allam, M., Phan, M.V.T., Abid, N., Rujeni, N., Abuzaid, N., Ismael, N., Elguindy, N., Top, N.M., Dia, N., Mabunda, N., Hsiao, N.-Y., Silochi, N.B., Francisco, N.M., Saasa, N., Bbosa, N., Murunga, N., Gumede, N., Wolter, N., Sitharam, N., Ndodo, N., Ajayi, N.A., Tordo, N., Mbhele, N., Razanajatovo, N.H., Iguosadolo, N., Mba, N., Kingsley, O.C., Sylvanus, O., Femi, O., Adewumi, O.M., Testimony, O., Ogunsanya, O.A., Fakayode, O., Ogah, O.E., Oludayo, O.-E., Faye, O., Smith-Lawrence, P., Ondoa, P., Combe, P., Nabisubi, P., Semanda, P., Oluniyi, P.E., Arnaldo, P., Quashie, P.K., Okokhere, P.O., Bejon, P., Dussart, P., Bester, P.A., Mbala, P.K., Kaleebu, P., Abechi, P., El-Shesheny, R., Joseph, R., Aziz, R.K., Essomba, R.G., Ayivor-Djanie, R., Njouom, R., Phillips, R.O., Gorman, R., Kingsley, R.A., Neto Rodrigues, R.M.D.E.S.A., Audu, R.A., Carr, R.A.A., Gargouri, S., Masmoudi, S., Bootsma, S., Sankhe, S., Mohamed, S.I., Femi, S., Mhalla, S., Hosch, S., Kassim, S.K., Metha, S., Trabelsi, S., Agwa, S.H., Mwangi, S.W., Doumbia, S., Makiala-Mandanda, S., Aryeetey, S., Ahmed, S.S., Ahmed, S.M., Elhamoumi, S., Moyo, S., Lutucuta, S., Gaseitsiwe, S., Jalloh, S., Andriamandimby, S.F., Oguntope, S., Grayo, S., Lekana-Douki, S., Prosolek, S., Ouangraoua, S., van Wyk, S., Schaffner, S.F., Kanyerezi, S., Ahuka-Mundeke, S., Rudder, S., Pillay, S., Nabadda, S., Behillil, S., Budiaki, S.L., van der Werf, S., Mashe, T., Mohale, T., Le-Viet, T., Velavan, T.P., Schindler, T., Maponga, T.G., Bedford, T., Anyaneji, U.J., Chinedu, U., Ramphal, U., George, U.E., Enouf, V., Nene, V., Gorova, V., Roshdy, W.H., Karim, W.A., Ampofo, W.K., Preiser, W., Choga, W.T., Ahmed, Y.A., Ramphal, Y., Bediako, Y., Naidoo, Y., Butera, Y., de Laurent, Z.R., Africa Pathogen Genomics Initiative (Africa PGI)‡, Ouma, A.E.O., von Gottberg, A., Githinji, G., Moeti, M., Tomori, O., Sabeti, P.C., Sall, A.A., Oyola, S.O., Tebeje, Y.K., Tessema, S.K., de Oliveira, T., Happi, C., Lessells, R., Nkengasong, J., Wilkinson, E., 2022. The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance. Science 378, eabq5358. https://doi.org/10.1126/science.abq5358

Van Noorden, R., 2021. Scientists call for fully open sharing of coronavirus genome data. Nature 590, 195–196. https://doi.org/10.1038/d41586-021-00305-7

Vavrek, D., Speroni, L., Curnow, K.J., Oberholzer, M., Moeder, V., Febbo, P.G., 2021. Genomic surveillance at scale is required to detect newly emerging strains at an early timepoint. https://doi.org/10.1101/2021.01.12.21249613

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., ’t Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Wilson, K., Neylon, C., Montgomery, L., Huang, C.-K. (Karl), Handcock, R.N., Roelofs, A., Hosking, R., Ozaygen, A., 2022. Global Diversity in Higher Education Workforces: Towards Openness. Open Library of Humanities 8. https://doi.org/10.16995/olh.4809

Wise, J., de Barron, A.G., Splendiani, A., Balali-Mood, B., Vasant, D., Little, E., Mellino, G., Harrow, I., Smith, I., Taubert, J., van Bochove, K., Romacker, M., Walgemoed, P., Jimenez, R.C., Winnenburg, R., Plasterer, T., Gupta, V., Hedley, V., 2019. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discovery Today 24, 933–938. https://doi.org/10.1016/j.drudis.2019.01.008

Yehudi, Y., Hughes-Noehrer, L., Goble, C., Jay, C., 2022. COVID-19: An exploration of consecutive systemic barriers to pathogen-related data sharing during a pandemic.