**Reconciling Data Actionability and Accountability in Global Health Research**

**Nathanael Sheehana\* and Sabina Leonellia**

**aEgenis, Centre for the Study of Life Sciences, Exeter University, University, Exeter, England**

Nathanael Sheehan ORCID 0000-0002-2779-0976

Sabina Leonelli OCRID 0000-0002-7815-6609

Correspondence concerning this article should be sent to ns651@exeter.ac.uk or s.leonelli@exeter.ac.uk. These joint first authors contributed equally to this work.

**Authors bios**: Nathanael Sheehan is a PhD student at the University of Exeter studying on the Environmental Intelligence centre for doctoral training. Sabina Leonelli is a Professor of Philosophy and History of Science at the University of Exeter and holds the Chair of Philosophy and History of Science and Technology at the Technical University of Munich where she will co-direct a new "Public Science Lab" and the Ethical Data Initiative.

***For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.***

***Word Count: 7302***

**Abstract**

All too often, the requirements for actionability and accountability of data infrastructures are conceptualised as incompatible and leading to a trade-off situation where increasing one will unavoidably decrease the other. Through a comparative analysis of two data infrastructures used to share genomic data about the SARS-COV-2 virus, we argue that making data actionable for knowledge development involves a commitment to ensuring that the data in question are representative of the phenomena being studied and accountable to data subjects and users. This in turn presupposes that: (1) enough data are contributed by a wide and diverse set of relevant sources; (2) mechanisms of feedback and inclusion are set up to ensure that data contributors can participate data governance and interpretation, thereby helping to adequately contextualise data; and (3) accountability extends to the ways in which data infrastructures are run, financed and positioned vis-à-vis the communities they are meant to serve. Such a model of data sharing can only work on the understanding that data do not need to be easily accessible to be actionable; rather, actionability depends on the responsiveness and accountability of data infrastructures, and the efforts invested in ensuring open communication among contributors.

**Keywords**: *Data Infrastructures, SARS-COV-2, Pandemic Data Science*

## Introduction

Since the start of the COVID-19 pandemic, a robust debate emerged over the meaning and significance of openly sharing genomic data on SARS-COV-2 variants. Many leading researchers strongly defended free sharing and access to such data as a crucial and necessary step towards better understanding the virus and devising effective public health interventions as quickly as necessary. At the same time, the issue of equity and accountability for such sharing was raised as a counterpoint to such unrestricted openness, with arguments made around the importance of ensuring that whatever was shared was of good enough quality, easily traceable to the original authors and appropriately crediting the researchers and data subjects involved in its creation. The redeployment of the Global Initiative on Sharing Avian Influenza Data (GISAID) in 2020 as a COVID-19 genomics sharing platform became a poster child for this debate, since its main selling point as a data infrastructure was precisely the promise of accountability in the form of credit due to the original data creators. As we discuss in detail below, the GISAID model was challenged by researchers concerned that data use within this infrastructure was too restricted and limiting to discovery, and in 2021 an open letter was published by the European Molecular Biology Lab and the European Bioinformatics Institute (EMBL-EBI) which demanded to make all COVID-19 related data accessible to one of their infrastructures without any restriction, focusing strongly on making such data immediately actionable through sophisticated labels, modelling tools and interoperability options. The tensions between these two data infrastructures, and what were widely portrayed as two different modes of data governance, intensified as concerns emerged over the funding and (lack of) responsiveness of GISAID leadership in 2022 and 2023.

Within this paper we are not interested in settling this debate or making pronouncements on which of these two data infrastructures was most effective in achieving their goals (an impossible task anyhow given how different these infrastructures are, as we illustrate in what follows and in previous work (Sheehan et al., 2024). Rather, we are interested in what this debate signals about broader trends and concerns in data science and related infrastructures. We take this episode, and the ongoing debates underpinning the data sharing efforts by both GISAID and EMBL-EBI, to exemplify a tension between *actionability* and *accountability* that is often found in data science initiatives within the biological and biomedical sciences, and particularly infrastructures aimed to serve transnational health research. On the one hand, engaging in data sharing efforts only makes sense if data can be used to support discovery, thereby becoming ‘actionable’ – a consideration that leads some advocates of open data to argue that any constraint on data circulation constitutes an obstacle to knowledge production, especially under emergency circumstances requiring a rapid response such as a pandemic. On the other hand, data infrastructures need to be accountable both to data contributors and to data users, thereby retaining the ability to credit those who contribute data and expertise and to track and justify what data are included (or excluded), how such data are processed, who is granted access and for which purposes. Accountability may thus involve restrictions on who may donate, access and re-use the data – and how. Hence all too often the requirements for actionability and accountability of data infrastructures are conceptualised as incompatible and leading to a trade-off situation where increasing one will unavoidably decrease the other, with severe implications either for the effectiveness and speed with which data can be deployed, or for the extent to which the provenance and processing of data can be responsibly documented.

Through a comparison of two platforms used to share genomic data about the SARS-COV-2 virus, the European Nucleotide Archive (ENA) – built by EMBL-EBI - which prioritises actionability and the GISAID which prioritises accountability, we argue that this is a false trade-off. Rather, we argue that the tensions between the two infrastructures stem from their different ways of pursuing data actionability, one of which focuses on immediate data use while the other prioritises data representativeness in the longer term. We show that it is not only possible, but necessary to reconcile these two approaches, thereby aiming towards an inclusive model of data sharing that enhances users’ ability to work with data at speed without sacrificing data protection measures and mechanisms of fair governance. Making data actionable for knowledge development involves a commitment to ensuring that the data in question are representative of the phenomena being studied. This in turn presupposes that: (1) enough data are contributed by a wide and diverse set of relevant sources; (2) mechanisms of feedback and inclusion are set up to ensure that data contributors can participate data governance and interpretation, thereby helping to adequately contextualise data; and (3) accountability extends to the ways in which data infrastructures are run, financed and positioned vis-à-vis the communities they are meant to serve. Crucially, such a model of data sharing can only work when accepting that data do not need to be transparent, or even easily accessible, in order to be actionable; and that while including a variety of contributors in efforts of data governance and interpretation may slow down the pace of discovery relative to having no governance at all, it is likely to boost the robustness and quality of outputs, as well as facilitate a more effective and faster use of data in the long term.

The paper is organised as follows. First, we briefly introduce the role of the ENA and GISAID in the context of SARS-CoV-2 data sharing and the how they came to represent commitments to actionability and accountability. We then utilise this case to probe the varying definitions of actionable and accountable data governance as embodied by these data infrastructures. Finally, we challenge the trade-off framing by arguing for the role of inclusive engagement in data governance for global health research.

1. **Global Health Research Infrastructures for SARS-CoV-2**

The dissemination of viral genetic data pertaining to SARS-CoV-2 is often celebrated as a seminal achievement in Open Science (Maxmen 2021; Besançon et al., 2021; Molldrem et al., 2021). Notably, the quest to track the spread and mutations of SARS-CoV-2 occasioned the most widespread sharing of viral data in history, be that predominantly through some kind of open access channel (Arita, 2021). This practice of openly sharing viral data can be traced back to decades of concerted institutional efforts and investments to facilitate the free, immediate and widespread sharing of sequence data (Leonelli & Ankeny, 2012; Hilgartner, 2017; Maxson Jones et al., 2018; Strasser, 2019). The large volume of multi-omics data disseminated on SARS-CoV-2 is linked to ongoing advances in the development of genomic surveillance infrastructures (Vavrek et al., 2021; Chen et al., 2022). These global data infrastructures are crucial to identifying possible novel threats to human health, while also shaping the tools, methods and analytics used by data scientists to research the virus and its impact on patients (Dupré & Leonelli, 2022). In earlier work (Sheehan et al., 2024), we identified GISAID and ENA - as key genomic surveillance infrastructures for SARS-CoV-2 with divergent models of data governance. These differences are apparent in the conflicting reuse policies offered by either infrastructure.

For GISAID, their data sharing model deviates from a straightforward open sharing paradigm, instead requiring users to authenticate their identity and obtain permission from data producers prior to republishing or linking GISAID genomes (Elbe & Buckland-Merrett, 2017). GISAID’s operationalisation of accountable data sharing for avian flu in 2006 rendered it a coherent progression to adapt its modus operandi in 2020 to encompass the newly emergent SARS-CoV-2 virus. Through building an early relationship with the Chinese Centre for Disease, Control and Protection (CDC), GISAID was among[[1]](#endnote-1) the first to publish a full genome of SARS-CoV-2 publicly. The dissemination of this data happened in the EpiCov data repository housed within GISAID, a dedicated SARS-CoV-2 genetic database built with the existing model of accountable – for the contributor and secondary user – terms of use. The EpiCoV database included an ecosystem of proprietary analytical tools including sequence alignments, diagnostic primers and probes, 3D models of proteins, phylogenetic trees, as well as specialised tools focused on domain specific areas such as wastewater management. In addition to this, GISAID further provides daily analysis figures from EpiCoV data covering lineage frequencies, emerging variants, global occurrences, a full genome tree, temporal distribution of lineages, representative phylogenetics of sequences and phylodynamics of variants of interest and under monitoring (Khare et al., 2021). With over 16 million assembled genomes sequenced by May 2024, GISAID is at the time of writing the premier open access data repository for SARS-CoV-2 (GISAID, 2024).

Conversely, the ENA allow bulk downloads of data without user tracking and for free, fostering linkage and cross-analysis with an array of data types encompassing viral sequences, host sequences, expression, protein and biomolecular data, microscopic images of the virus and research literature pertaining to the study of life science phenomena. This has been a core tenant of the ENA’s philosophy for the past three decades (Cameron, 1987, 1988) and has resulted in the archive accumulating petabytes of data in which nearly doubles year on year (Cochrane et al., 2012; Yuan et al., 2024). It was then no surprise that during the early stages of the SARS-CoV-2 pandemic the ENA would come to build a dedicated infrastructure for the collection, analysis and dissemination of SARS-CoV-2 sequence data. The new portal for SARS-CoV-2 was funded by the European Commission and was built and managed by the ENA and ELIXIR – a distributed infrastructure for life science information from over 250 research institutes in Europe (Stroe, 2020). The portal mission revolves around accelerating research through data sharing by curating data to enhance interoperability across various platforms. The academic article on the ENA portal (Harrison et al., 2021) elucidated the range of datasets, user tools, and the repository's alignment with the FAIR (Findable, Accessible, Interpretable, and Reusable) governance framework (Wilkinson et al., 2016), underscoring the criticality of open-access data dissemination at the fastest possible rate, thereby enabling analysis in real time. Although at the time of writing ENA has around 8 million less assembled sequences compared to GISAID, it primarily concentrates on publishing primarily raw viral data, so users can reassemble genomes based on annotations rather than taking the assembled genome as a complete ground truth. Moreover, services built by the ENA did not exist in isolation but was rather part of a broader array of services built by bioinformaticians at the EMBL-EBI during the Covid-19 pandemic to support the submission, analysis and dissemination of SARS-CoV-2 data (Rahman et al., 2024).

The differences in these two data models came to a point of escalation on the 29th of January 2021 the governing board of EMBL-EBI posted a public letter calling for a greater “openness” in sharing SARS-CoV-2 genome data (EMBL-EBI, 2021). The letter argued that “to unleash the fast flow of research advances” the scientific community must remove all formal barriers which restrict data sharing and share all SARS-CoV-2 genome sequences to one of a triad of state genomic surveillance programs who are members of the international nucleotide sequence database (INSDC) - European Nucleotide Archive, National Institute for Health of USA, and the DNA Data Bank of Japan (Stroe, 2021). The letter received 826 signatures (at the time of writing) and, building on these organisations’ long history of open data sharing, was promoted by two Nobel Laureates, numerous heads and directors of bioinformatic programs and many researchers at the cutting edge of genome sequencing (EMBL-EBI, 2021). The letter did not mention GISAID by name, yet it clearly targeted GISAID’s unkept promise to swiftly release protected genomes from their data repository into one that was completely open. The letter argued that infrastructures such as GISAID were limiting the extent to which data could be accessed, which may negatively affect the insight, pace and breadth of future COVID-19 research. As the pandemic unfolded, researchers began to criticise the quality (Gozashti & Corbett-Detig 2021) and pace (Kalia et al., 2021) of the GISAID data repository and demanded raw data that GISAID was not providing (Yehudi et al., 2022).

## 3. Interpretations of Actionability and Accountability

As signalled in our introduction, we are not primarily interested in adjudicating whether or not GISAID is less dependable vis-à-vis other efforts such as ENA. Rather, we wish to focus on a key tension between data actionability and accountability that this debate exemplifies, and in our view permeates data governance and infrastructure-building initiatives well beyond the COVID-19 genomics case. This is the tension between making data *actionable*, in the sense of making it possible for researchers to work with the data for a variety of purposes, and making data work *accountable*, thereby ensuring that data contributors and stewards are appropriately credited and held responsible for their choices and methods in creating and processing the data.

To better analyse this perceived tension, let us consider each characteristic in turn, starting from the idea of data actionability. Within the relevant literature, one finds a plurality of meanings and definitions for actionability. Actionability orbits around the opportunities for practical application of data, thereby underscoring the availability of analytic tools such as statistical modelling, algorithmic processing, and machine learning techniques (Chen et al., 2012; Kitchin, 2014; Lyon, 2014; Aradau & Blanke, 2015; Kitchin & McArdle, 2016;). Actionability typically necessitates the transformation of raw data into comprehensible quantitative presentations, such as graphs or tabular forms (Jasanoff, 2017), or data models (Leonelli, 2019; Bokulich & Parker, 2021). This is to say, actionability is contingent upon the structured depiction and contextualisation of data, with an underlying emphasis on the creation of strategies to make the re-use of data efficient and relevant to specific situations of inquiry (Baack, 2015; Erkkilä, 2016; Mulder et al., 2016). Data actionability is thereby dependent on the extent to which data can be shared among - and systematically handled by - research teams, the public, and/or institutions, particularly via expert-curated data repositories that can enhance the trustworthiness of the data and related visualisations / analytics tools (Tempini & Leonelli, 2021). Building on this scholarship, we take data actionability as the *capacity to re-purpose data for novel goals, methods of analysis and research settings,* which in turn depends on how data users interpret the prospective evidential value of the data.

Turning now to the concept of data accountability, we take this to indicate the degree of *responsiveness of data work and infrastructures to a wide variety of prospective users and situations* - and particularly the extent to which it is possible to identify and engage with individuals and groups who made decisions around the data (how they were created, for which purposes, what metadata were recorded and why, in which format... and so forth), which in turn endows such individuals and groups with a degree of responsibility around the characteristics and prospective useability of the data. Literature that focuses on data accountability typically entails a focus on ethics, fairness, and the potential for negative outcomes as integral components of research design or governance (Watson, 1996; Woolgar & Neyland, 2013; Garfinkel, 2023). Fox (2007) differentiates accountability in the realm of scientific data into two categories: "soft" accountability, wherein researchers must elucidate and justify any data inconsistencies within a relevant context or thereby being answerablefor their data, and "hard" accountability intensified by the possible consequences of non-adherence, such as expulsion from a data repository or forfeiture of funding grants. Drawing from this, Mayernik (2017) conceptualises accountability in open data as being in a linear relationship with transparency; where "hard" accountability ensures comprehensive adherence to all data management lifecycle requirements, while "soft" accountability entails mere online data publishing with limited metadata and responsiveness from the data provider (a problematic view on accountability to which we return in what follows).

Based on these conceptions of what data actionability and accountability may mean for data practitioners, it may be argued that, given their set-up and priorities, GISAID and ENA have come to exemplify two different strategies in data governance. ENA adopted as its core mission the capacity to re-purpose data quickly for novel goals and research settings, and thereby has come to prioritise data actionability by focusing less on concerns with data access and governance and more with issues of workability and interoperability of data, with a tremendous amount of work devoted to making data open to unexpected re-purposing through interoperable standards/tools. This emphasis on actionability has seemingly taken attention away from accountability understood as *responsiveness* to the social – rather than solely technical - conditions under which users may or may not be induced to participate in the work of the platform.

GISAID has instead strongly emphasised its commitment to accountability, interpreting this notion as tightly linked to the comprehensiveness and inclusivity of the platform: that is, the extent to which the data garnered by the platform constitute a representative sample of the populations under study, by virtue of including donations from as many countries as possible, irrespective of the knowledge of data stewardship and processing of data formats associated to specific labs and contributors. Especially in the case of SARS-COV-2 data sharing, where an immediate task is the identification of emerging variants of concern, it is critical to sample mutations from all over the globe to be able to spot new dangerous mutants as soon as they arise. Representativeness of samples and related data in this case cannot be enhanced or otherwise manipulated through computational tools (such as those used to correct for bias in other cases), but rather relates directly to the geographical locations of data sources and thus to providing incentives to share data to contributors from all around the world. Given these priorities, ensuring the representativeness of data samples through appropriate data governance and tech choices for GISAID meant acknowledging that science is not a level playing field and that working at the methodological cutting-edge does not necessarily support inclusion and trust within the global research community. This is a version of accountability that privileges responsiveness over transparency, while also noting the importance of fostering trustworthy data sharing as means to encourage more sharing across disparate users.

We think that there is much to learn from the version of accountability preferred by GISAID, especially when compared to understandings of accountability that are tightly linked to transparency as an unequivocally positive value. The latter interpretation places too little emphasis on the ability to respond to research situations in ways that take account of the implications of data use and of the conditions under which different users may or may not be able to engage with a given platform. Data scientists' ethical oversight and accountability are intertwined with technical and institutional decision-making, including the division of labour in data projects, the timing and type of expertise needed, the communication protocols between computer scientists and domain-specific experts, and the acknowledgment of all project contributors (Leonelli, 2016); *as well as* concerns with inclusivity and attention towards enabling a variety of different perspectives and contributions to interact with data platforms are part and parcel of those platforms’ accountability to their users (Leonelli, 2023). Transparency thus can be useful and sometimes even indispensable to enhance accountability, as Mayernick argued; but it needs to be supplemented with a social environment that fosters reflection around who, and in which capacity, is being encouraged and expected to be a user – and who may, instead, be excluded or unable to engage with a specific development, and with which implications. In other words, *accountability entails considerations of engagement:* who and what is being included in a data practice, for which purpose and with which implications.

## Challenging the trade-off framing: The role of inclusive engagement

With these ideas around accountability in mind, let us then look more closely at the putative trade-off between actionability and accountability that may be said to characterise the tensions between ENA and GISAID, and it is often invoked in broader discussions around different models of data governance. What does the idea of trade-off indicate? According to the *Cambridge Dictionary* (2024), a trade-off entails the “balancing of two opposing situations or qualities, both of which are desired”, and/or a “situation in which the achieving of something you want involves the loss of something else which is also desirable, but less so”.  In our view, thinking of actionability and accountability as a trade-off in this sense is unnecessary, grounded on a problematic understanding of Open Science as a whole and potentially damaging to research, since it can hamper data sharing efforts within the biological and biomedical sciences. Why this may be the case becomes clear when focusing on a common concern to both sides: data representativeness.

Making data actionable for knowledge development presupposes that the data are representative of the phenomena being studied, which in this case is the breadth and characteristics of the SARS-COV-2 variants emerging as the virus spreads and evolves. Data representativeness consists of the extent to which data can be taken to reliably represent the target population, which in the case of COVID-19 data depends on the degree to which a data infrastructure proves welcoming to contributors from all geographical locations affected by the pandemic. Accountability is thus core to the quest for a body of data which can be trusted by users to reliably represent global phenomena such as a pandemic. At the same time, providing transparent and responsive governance mechanisms for data sharing, together with a set of tools and standards that can facilitate data analysis, also constitutes an important incentive for prospective contributors to share data and trust a platform to substantively support global public health efforts. Actionability as interpreted by ENA is thus also an integral part of the quest for representative data, given its role in increasing users’ confidence that the data being disseminated will be processed and interpreted to facilitate novel discoveries in the future, and that it is therefore worth participating in the data sharing effort and consulting data infrastructures when planning new research. Accountability and actionability thus become complementary part of the same endeavour, which is to develop trustworthy, reliable, sustainable data infrastructures for long-term use. This is not necessarily in conflict with the requirement for speed which is so often cited as a prime motivation for unconstrained and unregulated data sharing, since while governance systems may slow down immediate access to the data, they need no limit that access and can actually improve efficiency in the longer term, for instance by fostering the use of metadata to track data provenance. As argued in the case of COVID-19 data sharing more broadly, fast science does not mean rushing research to the point of losing sight of its longer-term significance and requirements (Leonelli, 2020), nor believing that it can be value-free or neutral (El-Hani & Machado 2020).

For example, one could problematise ENA as exhibiting limited accountability of its governance structure by having no limits on how data are accessed or by whom, ENA have less foresight on potential reuse scenarios. Not only does this affect the trust of researchers and institutions who are concerned about their data being scooped or misused, it also gives the impression that the repository is uninterested in questions of sovereignty, equity and digital rights, particularly research happening in low resourced environments. GISAID’s data governance model can be problematised based on its limited metadata on the provenance and processing of data in the system, leading to researchers fearing that the data found therein would be unreliable or incomplete, and ultimately refraining from using them. This lack of attention paid to the features of the data and the conditions under which they can be analysed and re-used is also reflected in the limited types of data available – as predominantly focuses on assembled genomic sequence GISAID predominantly focuses on assembled genomic sequence. The narrow data scope not only hampers the comprehensive understanding and interpretation of the genomic data, but also constrains the ability of researchers to draw interdisciplinary connections and insights. Furthermore, the limited metadata provision may be argued to compromise the reproducibility and validity of research based on the GISAID database[[2]](#endnote-2) (Liu et al., 2023 and Crits-Christoph et al., 2023￼).

We therefore argue that enhancing the representativeness of data needs to be a core concern for data infrastructures facing public health, and that this involves three core requirements: (1) provide incentives to *diversify data contributions*, so that enough data are contributed by a wide and diverse set of relevant sources; (2) set up *mechanisms of feedback and inclusion* to ensure that data contributors can participate in data governance and interpretation no matter where they are based and which facilities they have access to; and (3) *conceptualise accountability* as extending beyond specific instances of data use, and especially to the ways in which data infrastructures are run, financed and positioned vis-à-vis the communities that they are meant to serve (see box 1).

**Box 1. Enhancing the Representativeness of Data**

****

**Enhancing the representativeness of data collection presupposes that:**

* 1. enough data are contributed by a wide and diverse set of relevant sources;
  2. mechanisms of feedback and inclusion are set up to ensure that data contributors can participate in data governance and interpretation, thereby helping to adequately contextualise data;
  3. accountability extends to the ways in which data infrastructures are run, financed and positioned vis-à-vis the communities they are meant to serve.

These requirements pick up on both ENA and GISAID intuitions about what matters to the governance of data infrastructures, thereby learning useful lessons from the comparison between the two. On the one hand, it is possible to improve the data governance of ENA to enhance trust, willingness to engage with the platform and thus ultimately representativeness, while at the same time facilitating actionability. For instance, requiring registration by users can help to keep track of data usage and create communities of users engaged in similar activities, benefitting from reciprocal feedback and exchange, and building up trust and connections along the way. Requiring citation of the database can help to track provenance of data and the various pathways involved in data modelling and enrichment, which in turn can help establish the reliability and quality of the data vis-à-vis a variety of purposes and contexts. On the other hand, it is also feasible to enhance the actionability of data stored by GISAID, taking advantage of the technical innovation and interoperability standards (metadata, linkage tools, versioning and history tracking) pushed by ENA. Investing in better governance for prospective usage, such as for instance modelled by data platforms such as the Secure Anonymised Information Linkage databank (Tempini & Leonelli, 2018), can also improve representativeness and participation in data interpretation, which in turn improves actionability. Especially given the concerns surrounding GISAID management at the time of writing, it is clear that a more transparent decision-making process for how data access is granted would greatly improve users’ trust in GISAID as a data steward and therefore their willingness to donate their data to the platform.

Overcoming the trade-off framing of the relationship between accountability and actionability entails recognising that data actionability is not just about technical opportunities (such as extensive linkage, standardization, interoperability, multi-scale modelling). Rather, and as also highlighted by Mayernik, data actionability also requires ‘soft’ skills and governance strategies to manage transdisciplinary expertise of relevance to data interpretation, as well as investment in trustworthy governance to increase participation and thereby representativeness of data, and a governance strategy that is responsive to and responsible towards contributors and users. Recognising the importance of wide-ranging, inclusive engagement opportunities for actionability can help address some of the problems plaguing the two data infrastructures as considered in the previous sections.

The difference in priorities and resulting tensions within these systems is by no means limited to this case. Efforts to apply data protection laws such as General Data Protection Regulation (GDPR) to datasets used for research have also encountered strong resistance, particularly in the form of concerns that limiting access and exploration of existing datasets would result in much diminished capacity to use those data as evidence to support novel insights and research directions. In the absence of a sophisticated governance structure, and the funding to maintain it, this has indeed come to pass: many research institutions (such as hospitals, for instance) have become more conservative over their data sharing policies, making it hard for researchers to be granted access and virtually impossible to simply ‘explore’ the data without a precise and pre-existing commitment to specific forms of re-use. This is a problem particularly for exploratory research based on data mining techniques, where access to the data is seen as a foundational requirement to be able to ‘look around’ and fish for surprising findings, promising correlations, and new hypotheses.

## Conclusion

In conclusion, the debate on the sharing of genomic data on SARS-CoV-2 variants underscores the complexities of reconciling actionability and accountability in data science, especially in the context of global health emergencies. This paper has examined the operational models of GISAID and ENA, revealing that the perceived trade-off between making data actionable and ensuring accountability is not just oversimplified but fundamentally misleading. Instead, the experiences of these platforms illustrate that a nuanced approach to data governance, which simultaneously advances both principles, is not only feasible but essential for effective public health response.

Both platforms, despite their differences, demonstrate the importance of integrating robust data management with flexible access to foster innovation and timely intervention in public health crises. The GISAID model, emphasising accountability, ensures that data contributors are credited, fostering trust and continued participation from diverse global communities. On the other hand, ENA's focus on making data immediately actionable underscores the need for rapid data availability to address urgent public health needs. Through this analysis, we have identified key strategies that can reconcile the two approaches: enhancing the inclusivity of data contributions, implementing dynamic governance mechanisms that adapt to evolving research and public health landscapes, and ensuring that data infrastructures are responsive to both immediate and long-term research needs. By embracing these strategies, data infrastructures can overcome the false dichotomy between actionability and accountability, ultimately supporting more robust, equitable, and effective global health responses.

It is our hope that this paper's insights into the operational tensions between GISAID and ENA may serve to inform future data sharing initiatives. We conclude by advocating for a model of data governance that not only accelerates scientific discovery and global public health interventions but also respects and upholds the principles of equity and transparency in research practices. As the landscape of global health research continues to evolve, and further crises unfold, the lessons drawn from this debate will undoubtedly influence the design and implementation of more resilient and responsive data sharing systems for planetary health.

**Acknowledgement**

We thank colleagues at the European Bioinformatics Institute and the “Philosophy of Open Science for Diverse Research Environments” (PHIL\_OS) project for their feedback and support in this research. This paper reflects only the authors’ views and the Commission / Agency is not responsible for any use that may be made of the information it contains.

**Disclosure Statement**

No potential conflict of interest was reported by the author(s).

**Funding**

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 101001145).

## References

Aradau, C., & Blanke, T. (2015). The (Big) Data-security assemblage: Knowledge and critique. Big Data & Society, 2(2), 2053951715609066. https://doi.org/10.1177/2053951715609066

Arita, M. (2021). Open Access and Data Sharing of Nucleotide Sequence Data. 20(1), 28. https://doi.org/10.5334/dsj-2021-028

Baack, S. (2015). Datafication and empowerment: How the open data movement re-articulates notions of democracy, participation, and journalism. Big Data & Society, 2(2), 2053951715594634. https://doi.org/10.1177/2053951715594634

Besançon, L., Peiffer-Smadja, N., Segalas, C., Jiang, H., Masuzzo, P., Smout, C., Billy, E., Deforet, M., & Leyrat, C. (2021). Open science saves lives: lessons from the COVID-19 pandemic. BMC Medical Research Methodology, 21(1), 117. https://doi.org/10.1186/s12874-021-01304-y

Bokulich, A., & Parker, W. (2021). Data models, representation and adequacy-for-purpose. European Journal for Philosophy of Science, 11(1), 31. https://doi.org/10.1007/s13194-020-00345-2

Burki, T. (2023). First shared SARS-CoV-2 genome: GISAID vs virological.org. The Lancet. Microbe, 4(6), e395. https://doi.org/10.1016/S2666-5247(23)00133-7

Cambridge Dictionary, E. (2024, May 15). tradeoff. https://dictionary.cambridge.org/us/dictionary/english/tradeoff

Cameron, G. (1987). Graham Cameron material - EMBL Archive. EMBL-EBI Archive. https://archive.embl.org/index.php/graham-cameron-fonds

Cameron, G. N. (1988). The EMBL data library. Nucleic Acids Research, 16(5), 1865–1867. https://doi.org/10.1093/nar/16.5.1865

Chen, Chiang, & Storey. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36(4), 1165. https://doi.org/10.2307/41703503

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. Nature Genetics, 54(4), 499–507. https://doi.org/10.1038/s41588-022-01033-y

Cochrane, G., Alako, B., Amid, C., Bower, L., Cerdeño-Tárraga, A., Cleland, I., Gibson, R., Goodgame, N., Jang, M., Kay, S., Leinonen, R., Lin, X., Lopez, R., McWilliam, H., Oisel, A., Pakseresht, N., Pallreddy, S., Park, Y., Plaister, S., … Zalunin, V. (2012). Facing growth in the European Nucleotide Archive. Nucleic Acids Research, 41(D1), D30–D35. https://doi.org/10.1093/nar/gks1175

Crits-Christoph, A., Gangavarapu, K., Pekar, J. E., Moshiri, N., Singh, R., Levy, J. I., Goldstein, S. A., Suchard, M. A., Popescu, S., Robertson, D. L., Lemey, P., Wertheim, J. O., Garry, R. F., Rasmussen, A. L., Andersen, K. G., Holmes, E. C., Rambaut, A., Worobey, M., & Débarre, F. (2023). Genetic evidence of susceptible wildlife in SARS-CoV-2 positive samples at the Huanan Wholesale Seafood Market, Wuhan: Analysis and interpretation of data released by the Chinese Center for Disease Control. Zenodo. https://doi.org/10.5281/zenodo.7754299

Dupré, J., & Leonelli, S. (2022). Process epistemology in the COVID-19 era: rethinking the research process to avoid dangerous forms of reification. European Journal for Philosophy of Science, 12(1), 20. https://doi.org/10.1007/s13194-022-00450-4

Elbe, S., & Buckland‐Merrett, G. (2017). Data, disease and diplomacy: GISAID’s innovative contribution to global health. Global Challenges, 1(1), 33–46. https://doi.org/10.1002/gch2.1018

El-Hani, C. N., & Machado, V. (2020). COVID-19: The need of an integrated and critical view. Ethnobiology and Conservation, 9. https://ethnobioconservation.com/index.php/ebc/article/view/408

EMBL-EBI. (2021). Sign the letter to support data sharing for COVID-19. https://www.covid19dataportal.org/support-data-sharing-covid19/sign

Erkkilä, T. (2016). Global Governance Indices as Policy Instruments: Actionability, Transparency and Comparative Policy Analysis. Journal of Comparative Policy Analysis: Research and Practice, 18(4), 382–402. https://doi.org/10.1080/13876988.2015.1023052

Farrar, J., & Ahuja, A. (2022). Spike: The Virus v The People. The Inside Story. Profile Books. https://royalsociety.org/medals-and-prizes/science-book-prize/books/2022/spike/

Fox, J. (2007). The Uncertain Relationship between Transparency and Accountability. Development in Practice, 17(4/5), 663–671. https://www.jstor.org/stable/25548267

Garfinkel, H. (2023). Studies in Ethnomethodology. In Social Theory Re-Wired (3rd ed.). Routledge.

GISAID. (2024). GISAID. https://doi.org/10.17616/R3Q59F

Gozashti, L., & Corbett-Detig, R. (2021). Shortcomings of SARS-CoV-2 genomic metadata. BMC Research Notes, 14(1), 189. https://doi.org/10.1186/s13104-021-05605-9

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., … Apweiler, R. (2021). The COVID-19 Data Portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. Nucleic Acids Research, 49(W1), W619–W623. https://doi.org/10.1093/nar/gkab417

Hilgartner, S. (2017). Reordering Life: Knowledge and Control in the Genomics Revolution. The MIT Press. https://www.jstor.org/stable/j.ctt1qd8zf1

Jasanoff, S. (2017). Virtual, visible, and actionable: Data assemblages and the sightlines of justice. Big Data & Society, 4(2), 2053951717724477. https://doi.org/10.1177/2053951717724477

Kalia, K., Saberwal, G., & Sharma, G. (2021). The lag in SARS-CoV-2 genome submissions to GISAID. Nature Biotechnology, 39(9), 1058–1060. https://doi.org/10.1038/s41587-021-01040-0

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 Weekly, 3(49), 1049–1051. https://doi.org/10.46234/ccdcw2021.255

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 2053951714528481.

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society, 3(1), 2053951716631130. https://doi.org/10.1177/2053951716631130

Leonelli, S. (2016). Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2083), 20160122. https://doi.org/10.1098/rsta.2016.0122

Leonelli, S. (2019). What distinguishes data from models? European Journal for Philosophy of Science, 9(2), 22. https://doi.org/10.1007/s13194-018-0246-0

Leonelli, S. (2020). Data Science in Times of Pan(dem)ic. Harvard Data Science Review, 3(1). https://doi.org/10.1162/99608f92.fbb1bdd6

Leonelli, S. (2023). Philosophy of Open Science. Elements in the Philosophy of Science. https://doi.org/10.1017/9781009416368

Leonelli, S., & Ankeny, R. A. (2012). Re-thinking organisms: The impact of databases on model organism biology. Studies in History and Philosophy of Biological and Biomedical Sciences, 43(1), 29–36. https://doi.org/10.1016/j.shpsc.2011.10.003

Liu, W. J., Liu, P., Lei, W., Jia, Z., He, X., Shi, W., Tan, Y., Zou, S., Wong, G., Wang, J., Wang, F., Wang, G., Qin, K., Gao, R., Zhang, J., Li, M., Xiao, W., Guo, Y., Xu, Z., … Wu, G. (2023). Surveillance of SARS-CoV-2 at the Huanan Seafood Market. Nature, 1–3. https://doi.org/10.1038/s41586-023-06043-2

Lyon, D. (2014). Surveillance, Snowden, and Big Data: Capacities, consequences, critique. Big Data & Society, 1(2), 2053951714541861. https://doi.org/10.1177/2053951714541861

Maxmen, A. (2021). One million coronavirus sequences: popular genome site hits mega milestone. Nature, 593(7857), 21–21. https://doi.org/10.1038/d41586-021-01069-w

Maxson Jones, K., Ankeny, R. A., & Cook-Deegan, R. (2018). The Bermuda Triangle: The Pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project. Journal of the History of Biology, 51(4), 693–805. https://doi.org/10.1007/s10739-018-9538-7

Mayernik, M. S. (2017). Open data: Accountability and transparency. Big Data & Society, 4(2), 2053951717718853. https://doi.org/10.1177/2053951717718853

Molldrem, S., Hussain, M. I., & Smith, A. K. J. (2021). Open science, COVID-19, and the news: Exploring controversies in the circulation of early SARS-CoV-2 genomic epidemiology research. Global Public Health, 16(8–9), 1468–1481. https://doi.org/10.1080/17441692.2021.1896766

Mulder, F., Ferguson, J., Groenewegen, P., Boersma, K., & Wolbers, J. (2016). Questioning Big Data: Crowdsourcing crisis data towards an inclusive humanitarian response. Big Data & Society, 3(2), 2053951716662054. https://doi.org/10.1177/2053951716662054

Rahman, N., O’Cathail, C., Zyoud, A., Sokolov, A., Oude Munnink, B., Grüning, B., Cummins, C., Amid, C., Nieuwenhuijse, D. F., Visontai, D., Yuan, D. Y., Gupta, D., Prasad, D. K., Gulyás, G. M., Rinck, G., McKinnon, J., Rajan, J., Knaggs, J., Skiby, J. E., … Cochrane, G. (2024). Mobilisation and analyses of publicly available SARS-CoV-2 data for pandemic responses. Microbial Genomics, 10(2), 001188. https://doi.org/10.1099/mgen.0.001188

Sheehan, N., Botta, F., & Leonelli, S. (2024). Unrestricted Versus Regulated Open Data Governance: A Bibliometric Comparison of SARS-CoV-2 Nucleotide Sequence Databases. Data Science Journal, 23(1). https://doi.org/10.5334/dsj-2024-029

Strasser, B. J. (2019). Collecting Experiments: Making Big Data Biology. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/C/bo38870755.html

Stroe, O. (2020, February 24). Coronavirus data resources at EMBL-EBI. https://www.ebi.ac.uk/about/news/announcements/coronavirus-data/

Stroe, O. (2021, February 4). Show your support for open COVID-19 data. https://www.ebi.ac.uk/about/news/perspectives/open-data-letter/

Tempini, N., & Leonelli, S. (2018). Concealment and discovery: The role of information security in biomedical data re-use. Social Studies of Science, 48(5), 663–690. https://doi.org/10.1177/0306312718804875

Tempini, N., & Leonelli, S. (2021). Actionable data for precision oncology: Framing trustworthy evidence for exploratory research and clinical diagnostics. Social Science & Medicine, 272, 113760. https://doi.org/10.1016/j.socscimed.2021.113760

Vavrek, D., Speroni, L., Curnow, K., Oberholzer, M., Moeder, V., & Febbo, P. (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

Watson, G. (1996). Two Faces of Responsibility. Philosophical Topics, 24(2), 227–248. https://doi.org/10.5840/philtopics199624222

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., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

Woolgar, S., & Neyland, D. (2013). Mundane Governance: Ontology and Accountability. OUP Oxford.

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 (No. arXiv:2205.12098). arXiv. http://arxiv.org/abs/2205.12098

Yuan, D., Ahamed, A., Burgin, J., Cummins, C., Devraj, R., Gueye, K., Gupta, D., Gupta, V., Haseeb, M., Ihsan, M., Ivanov, E., Jayathilaka, S., Kadhirvelu, V. B., Kumar, M., Lathi, A., Leinonen, R., McKinnon, J., Meszaros, L., O’Cathail, C., … Cochrane, G. (2024). The European Nucleotide Archive in 2023. Nucleic Acids Research, 52(D1), D92–D97. https://doi.org/10.1093/nar/gkad1067

Author Notes

1. Although GISAID claims to have released the first whole genome sequence of SARS-CoV-2 alongside the China CDC, this is somewhat disputed in the virology community. Edward Holmes claims a pivotal role in the narrative of SARS-CoV-2: alongside his colleague, Zhang Yong-Zhen, who claim to be the first to publish the initial SARS-CoV-2 genome (Burki 2023). However, instead GISAID, Holmes chose to disseminate this significant discovery via virology.org. According to Holmes and Zhang, this event transpired a full twelve hours in before the official release of the genome by both GISAID and the Chinese authorities (Farrar and Ahuka 2022). [↑](#endnote-ref-1)
2. Notably, an intellectual contention surfaced post-outbreak within the virological community, as elucidated by the juxtaposition of the Liu et al and Crits-Christoph et al studies. Liu et al's data, originally disseminated on the GISAID platform, was subsequently used by Crits-Christoph et al in their analysis bolstering the early wet market transmission hypotheses. However, this data was later withdrawn from GISAID, prompting an inquiry into whether Crits-Christoph et al might have contravened the infrastructures terms of use. Adding a layer of complexity, Crits-Christoph et al disseminated a preprint of their findings prior to the official publication by the Chinese Center for Disease Control. Both manuscripts, through environmental sampling and sequencing, lean towards the hypothesis of a super-spreader event at the wet market. Yet, an undercurrent of contention exists; Crits-Christoph et al have suggested that the Chinese Center's ambition to publish in a high impact journal – such as Nature - might have delayed data release, potentially hindering the collaborative endeavours of the wider virology community. [↑](#endnote-ref-2)