From portal to publication: Searching for epistemic diverity in genomic surveillance infrastructure

Nathanael Sheehan, Sabina Leonelli & Federico Botta

2022-12-14

Background

The recent outbreak of SARS-CoV-2 has highlighted the crucial importance of timely sharing of genomic data and metadata for public health purposes. This renewed focus on scientific data sharing has sparked a debate between the Global Initiative on Sharing Avian Influenza Data (GISAID) and the Covid-19 Data Portal over what constitutes responsible viral data sharing in the context of emergency science.GISAID, which was founded in 2008 to monitor global influenza outbreaks, requires users to authenticate their identity and agree not to republish or link GISAID genomes with other datasets without permission from the data producer. The Covid-19 Data Portal, on the other hand, is an infrastructure launched by the European Commission in April 2020 to share - without restrictions - scientific data, including biological protein, expression, networks, imaging data, academic literature and socio-economic data. To date, little research has been done to evaluate the effectiveness of each infrastructure's open data policy in supporting epistemic diversity. Our aim is to address this gap by conducting a quantitative analysis of the two infrastructures in terms of their support for epistemic diversity, operationalised as the number of authors, institutions, ontologies, and geographies represented in global publications that make use of data from either infrastructure.

Methods

We accessed publications mentioning either infrastructure using the dimensions analytics api between between January 2020 and October 2022. The returned data underwent a scientific mapping using bibliographic methods from the `biblometrix` R package, including: general summary statistics, collaboration networks (authors, institutions, countries), co-citation Networks (authors, references, journals), coupling networks (references,authors,sources,countries), co-occurrences networks (authors,journal,keyword,title,abstract). We use features extracted from the networks to study the spatio-temporal heterogeneity in global scientific production of publications and submissions to SARS-CoV-2 genomic surveillance data portals.

Results

In our search, we found 20,955 publications in total - 11,256 in GISAID and 9,690 in the Covid-19 Data Portal. Both databases experienced a surge in mentions during the 2020-2021 period, but displayed differing annual growth rates, with GISAID being significantly more productive at 25.7% compared to the Covid-19 Data Portal's -5.7%. Based on our estimates, the Covid-19 Data Portal has an international co-authorship rate of 78.71%, with the United Kingdom (0.1342), United States (0.1223), Germany (0.1081), China (0.0650), and France (0.0580) being the most productive countries in terms of frequency. GISAID has a slightly lower international co-authorship rate of 68.3%, with the United States (0.2147), China (0.1326), India (0.0613), United Kingdom (0.0564), and Italy (0.0449) being the most productive countries.Given the short time span of publication for both databases, the number of preprint publications is relatively low, at 3.1% for GISAID and 5.2% for the Covid-19 Data Portal.Our results show that there is a significant difference in the way each

open data infrastructure supports epistemic diversity in terms of collaboration over time and space. GISAID is more likely to support forms of epistemic diversity based on geographical plurality, particularly among low and middle-income countries as well as high-income countries. In contrast, the Covid-19 Data Portal is more likely to support forms of epistemic diversity based on disciplinary expertise and experimental research methods.

Conclusion

In conclusion, our findings suggest that both open data platforms can support various forms of epistemic diversity in global scientific production, and that they should work together rather than against each other. Our study highlights the inadequacy of the normative framing of open data as a dichotomy between open and closed, and calls for a more situated understanding of open data in order to facilitate the rapid and effective sharing of scientific knowledge during times of crisis. This work aims to challenge existing definitions of open data and offer new insights into how open data can support epistemic diversity in the fields of genomic science and data science as a whole.

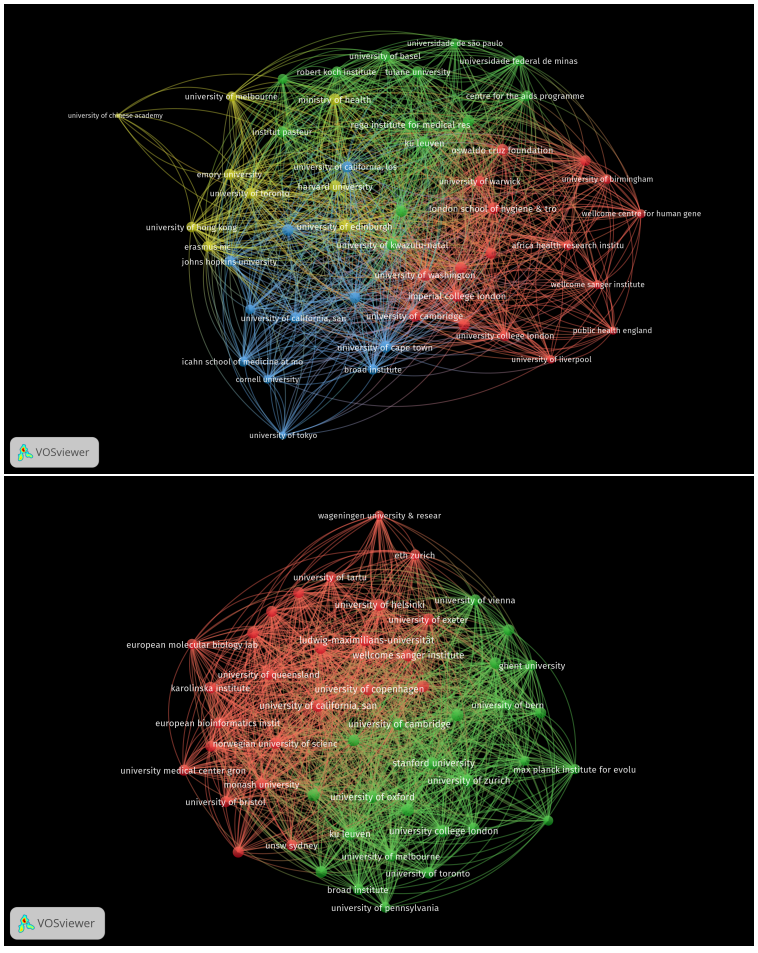

Figure 1: Above: Institutions who collaborate together using GISAID data (Size: 5469; Density: 0.005; Diameter: 9; Distance: 0.13; Avergae Path: 3.26).Below: Institutions who collaborate together using the covid-19 data portal (Size: 5786; Density: 0.33; Diameter: 8; Distance: 0.22; Average Path: 2.85')