Diverse enough? A systematic review of genome data portals during the sars-cov-2 pandemic

Nathanael Sheehan, Sabina Leonelli & Federico Botta

2022-12-09

Abstract

Background

Open data too has undergone a pandemic. The two common strains of open data - gratis or libre - are often described as antithetical to one another. Such a framing of open data has caused a fierce demarcation debate on what constitutes as responsible data sharing during a time of emergency science. This debate has largely played out between GISAID and the Covid-19 data portal in publications, public letters and leaked emails, yet little work has been done to quantitatively study the effectiveness of each platforms open data policy. We aim to address this research gap by conducting a quantitative analysis of the two platforms with regards to their support for epistemic diversity. We operationalise epistemic diversity as the number of authors, institutions, ontologies and geographies that are represented in global publications that make use of data from either platform.

Methods

20,955 publications (11,256 for GISAID and 9,699 for the Covid-19 Data portal) were accessed from 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). Using data generated from the networks, we perform a series of correlation tests to investigate the spatio-temporal heterogeneity in global scientific production of publications and submissions to SARS-CoV-2 genomic surveillance data portals.

Results

Our results show that there is a significant difference in the way gratis and libre open data platforms support epistemic diversity. Gratis open data platforms such as GISAID are more likely to support forms of epistemic diversity that are based on geographical location and institutional affiliation. In contrast, libre open data platforms such as the Covid-19 Data Portal are more likely to support forms of epistemic diversity that are based on disciplinary expertise and research methods.

Conclusion

We conclude that the two strains of open data are not antithetical to one another, but complementary. We show that the two strains of open data have produced different forms of epistemic diversity through global publications and recommend

Supplementary material - including code, data and presentations - can be accessed through the papers github repository.

*

Diverse enough: A scientific mapping of sars-cov-2 genome databases

Background

Open Data is often lumped into two categories: "gratis" or "libre". These two terms have famously defined a period of open source history where computer scientists fought over the semantic import of describing technical processes and artefacts. As the quote famously goes "free as in speach, not beer" libre is Stallmans preposition in articulating the Free Software Movement (FOSS) and his position in defending freedom as a first principle in software and data exchange.

The theme of epistemic diveristy has been explored in Open Science initiatives by identifying the number of factors in which can shape a diverse Open Science project. These include, but are not limited to a diversification of methods, characteristics of researchers, funding, geo-political location and intellectual property regimes (Leonelli 2021).One could say there are two main arguments for epistemic diversity; one is practical and the other is normative. The practical argument for epistemic diversity rests on the idea that diverse perspectives lead to new insights and theories. This argument has been put forward by many social epistemologists, including Helen Longino (2002) and Evelyn Fox Keller (1985). Longino argues that epistemic diversity is important because it leads to "the cognitive benefits of confrontation with alternative points of view" (2002, p. 74). In other words, when different perspectives are brought into contact with each other, they can challenge and improve each other. This process of confrontation and exchange is essential for scientific progress. Keller makes a similar point when she argues that "diversity... is an indispensable condition of fruitful scientific research" (1985, p. 21). She goes on to say that "without... diversity... there could be no growth in our knowledge" (1985, p. 21). Thus, according to the practical argument, epistemic diversity is important because it leads to better science. On the other hand, the normative argument for epistemic diversity rests on the idea that diverse perspectives should be included in science for moral or political reasons. This argument has been put forward by many feminist philosophers of science, including Sandra Harding (1986) and Donna Haraway (1991). Harding argues that we should strive for an "optimal mix" of voices in science because this would lead to "a more just society". These arguments are not mutually exclusive, and both can be used to support the case for epistemic diversity.

What has been the debate between the two portals?

Nature letter Gobels phd paper Global disparities paper Our previous work

How does this study fill the lacuna in research?

Data and Methods

The following section outlines the data and methods used as an emperical tool in this investigation. Although one of the authors have aforementioned in previous work against one of the methodologies used in this paper - biblometric analysis (Leonelli REF) - a degree of explanation may be needed.

Data collection

Include queries, where teh data was from, what catefories were searched against, time, language, article type and manual filtering efforts.

Data Analysis

- Explain scientific mapping, Bibliometric analysis, regression and state all the software and packages being used.
- Summary statistics
- Networks of publications
- Spatio-temporal correlations

Results

GISAID

Description	Results
MAIN INFORMATION ABOUT DATA	
Timespan	2020:2023
Sources (Journals, Books, etc)	1375
Documents	9699
Annual Growth Rate % -	85.86
Document Average Age	1.03
Average citations per doc	8.813
Average citations per year per doc	3.742
References	1
DOCUMENT TYPES	
article	7538
chapter	194
edited book	67
monograph	12
preprint	1880
proceeding	8
DOCUMENT CONTENTS	
Keywords Plus (ID)	6161
Author's Keywords (DE)	6161
AUTHORS	
Authors	56434
Author Appearances	101226
Authors of single-authored docs	136
AUTHORS COLLABORATION	
Single-authored docs	163
Documents per Author	0.172
Co-Authors per Doc	10.4
International co-authorships %	78.68

The Covid-19 Data Portal

Summary Statistics

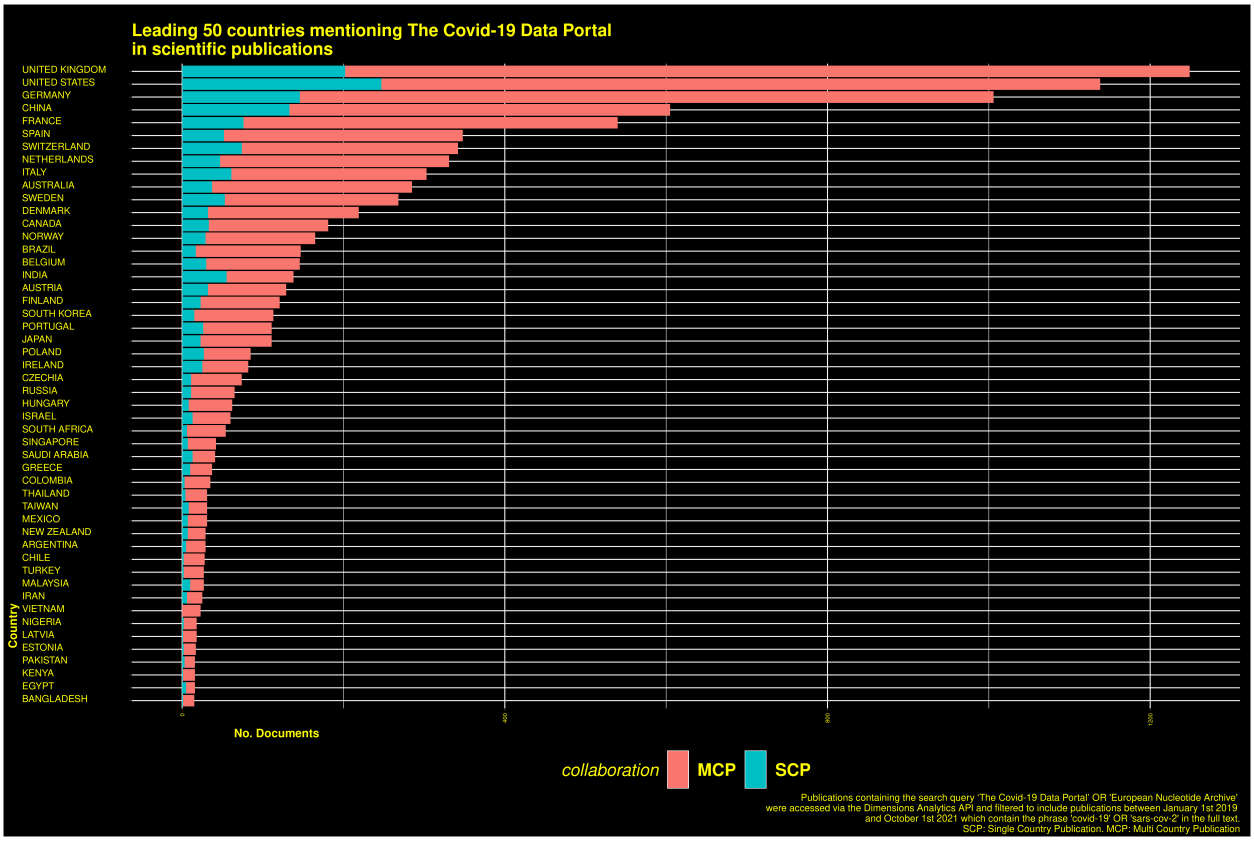

Figure 1: Data Governance - Open Science - Covid-19 - Epistemic Diversity

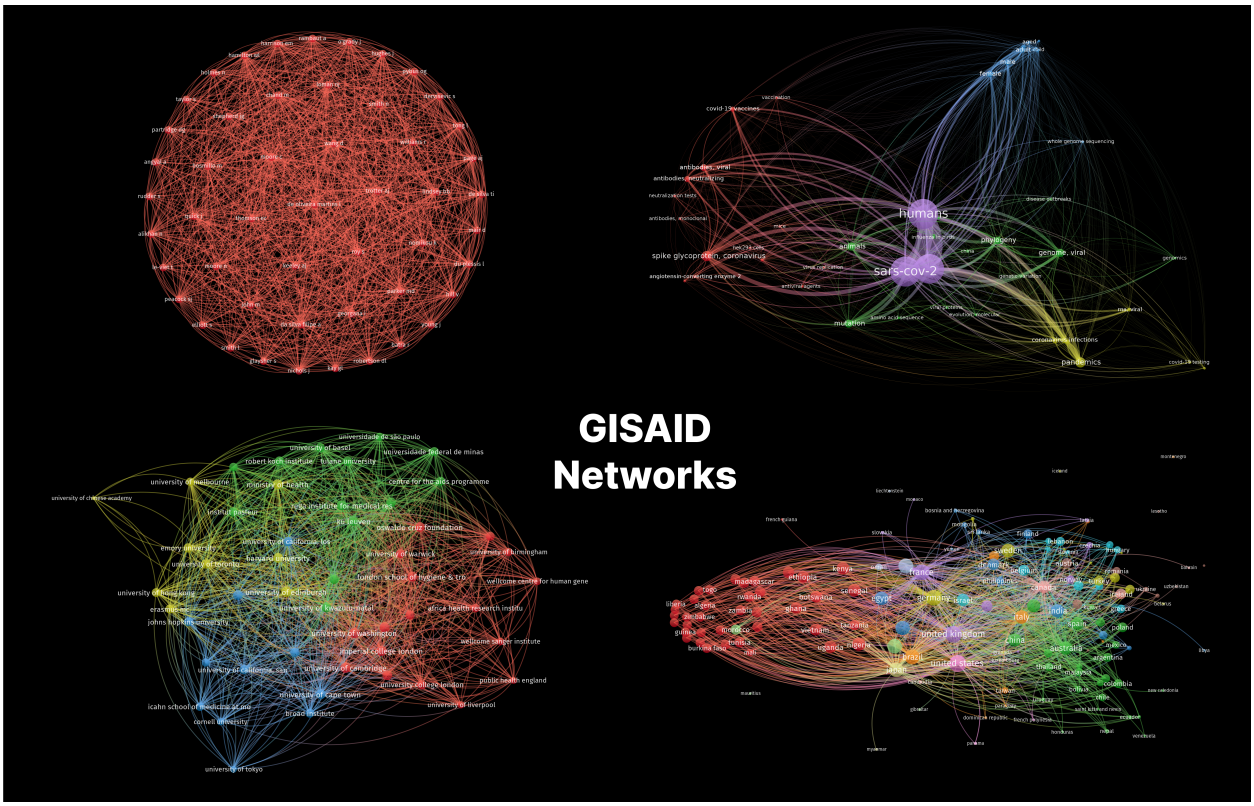

Figure 2: Data Governance - Open Science - Covid-19 - Epistemic Diversity

Description	Results
MAIN INFORMATION ABOUT DATA	
Timespan	2020:2023
Sources (Journals, Books, etc)	1375
Documents	9699
Annual Growth Rate % -	85.86
Document Average Age	1.03
Average citations per doc	8.813
Average citations per year per doc	3.742
References	1
DOCUMENT TYPES	
article	7538
chapter	194
edited book	67
monograph	12
preprint	1880
proceeding	8
DOCUMENT CONTENTS	
Keywords Plus (ID)	6161
Author's Keywords (DE)	6161
AUTHORS	
Authors	56434
Author Appearances	101226
Authors of single-authored docs	136
AUTHORS COLLABORATION	

Description	Results
Single-authored docs	163
Documents per Author	0.172
Co-Authors per Doc	10.4
International co-authorships %	78.68

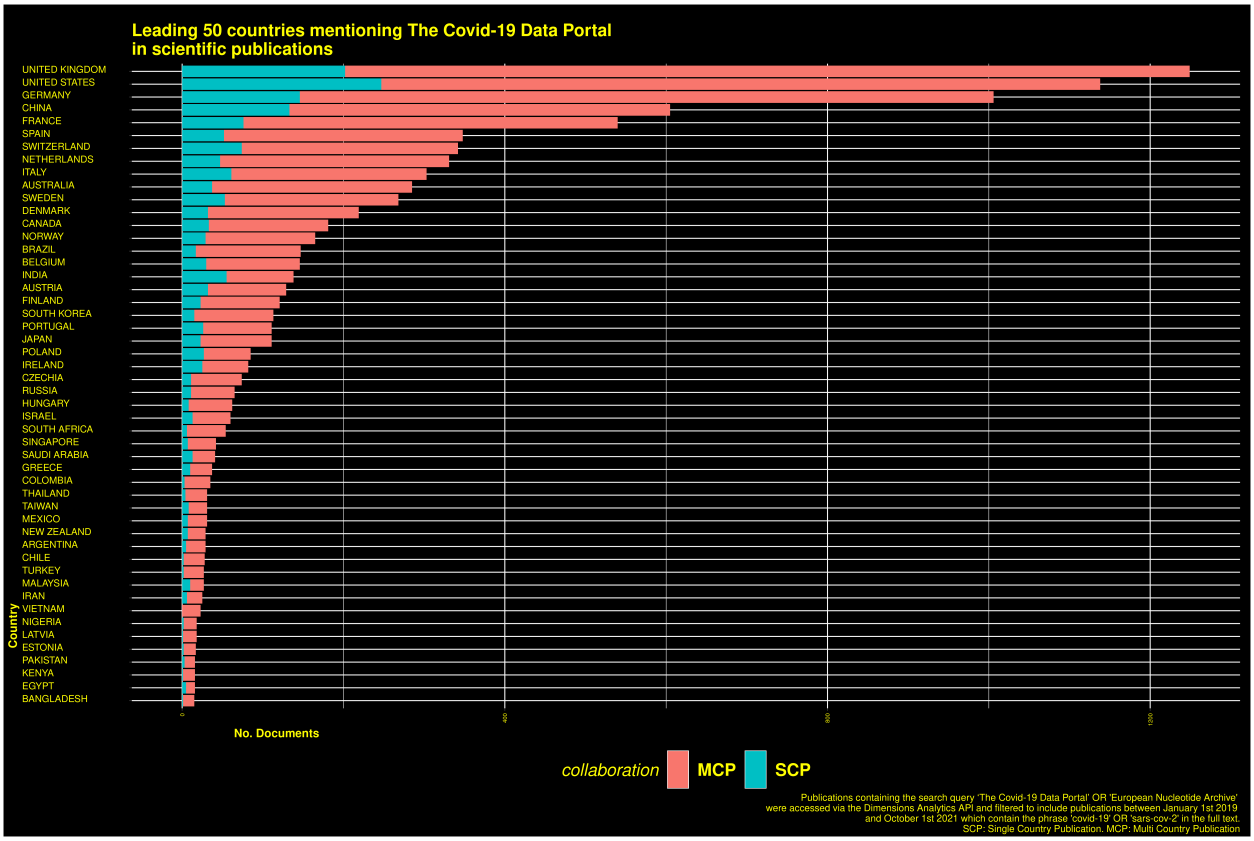

Figure 3: Data Governance - Open Science - Covid-19 - Epistemic Diversity

Networks

Discussion

Conclusion

References

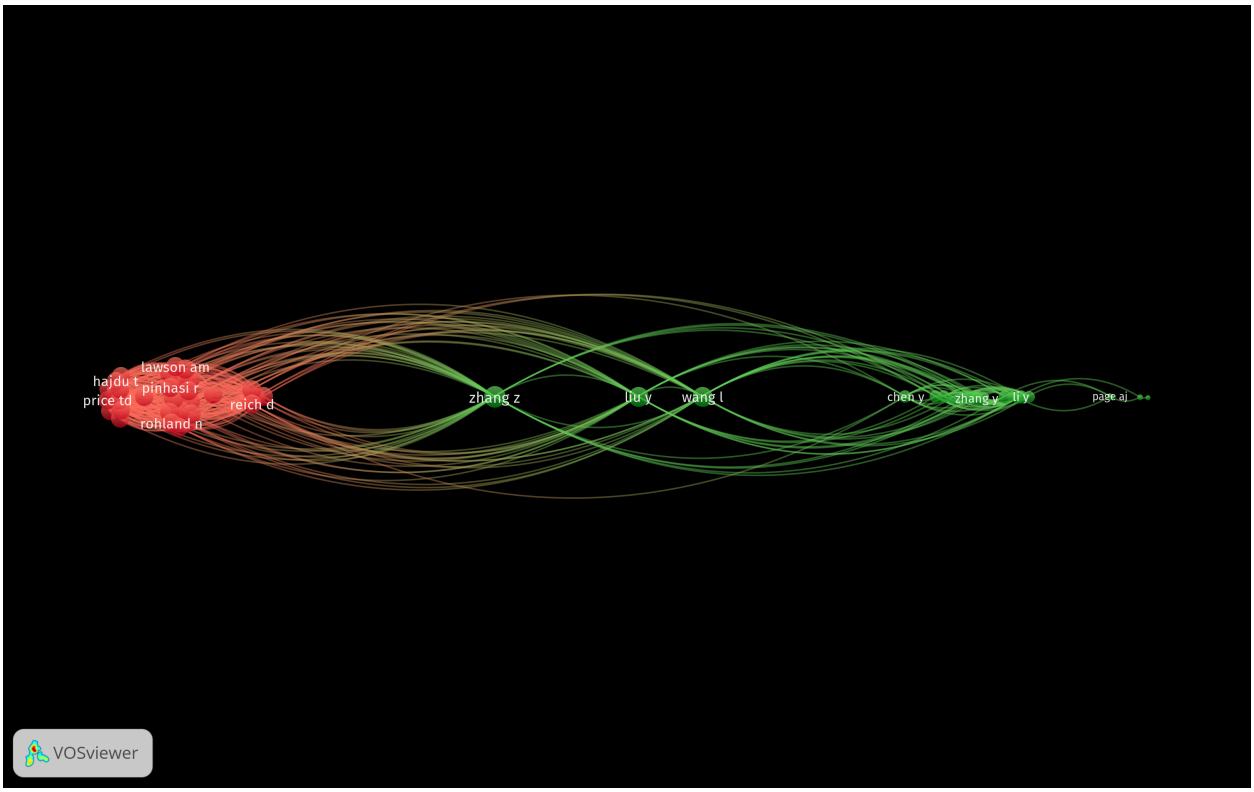

Figure 4: Data Governance - Open Science - Covid-19 - Epistemic Diversity

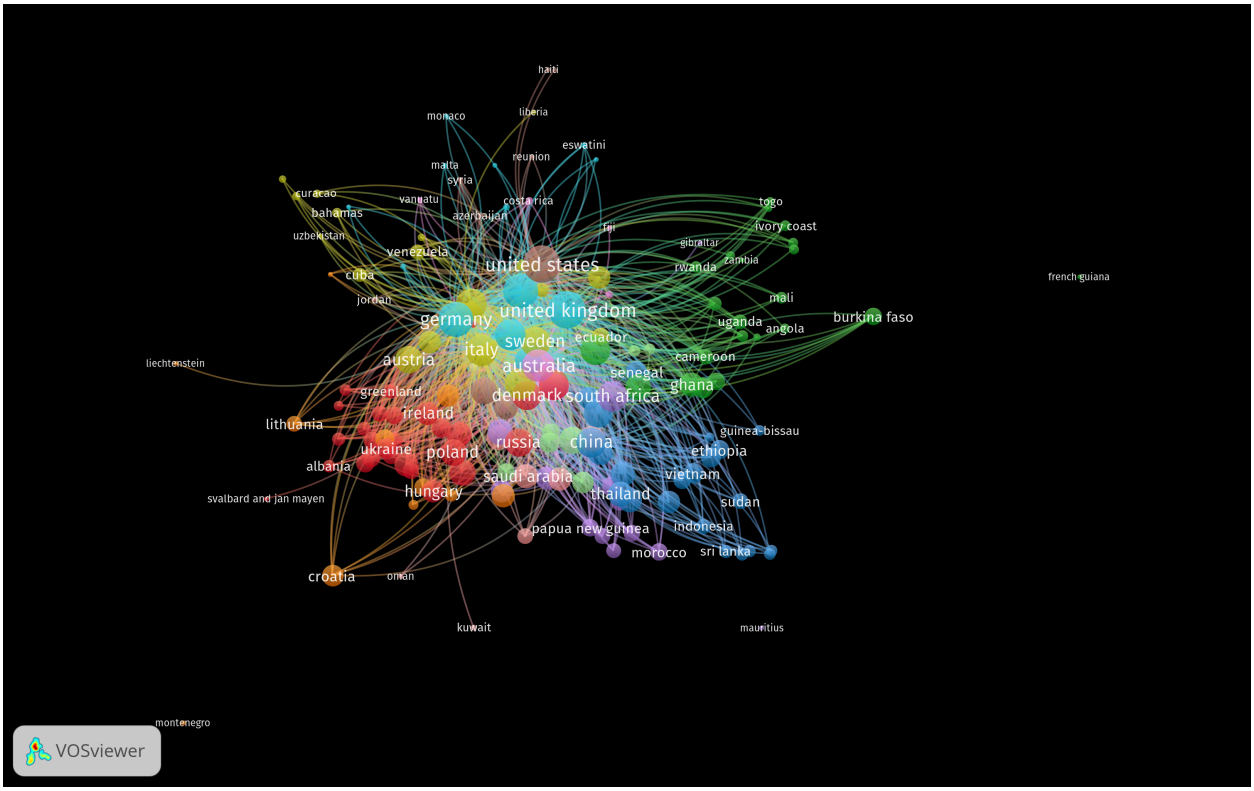

Figure 5: Data Governance - Open Science - Covid-19 - Epistemic Diversity

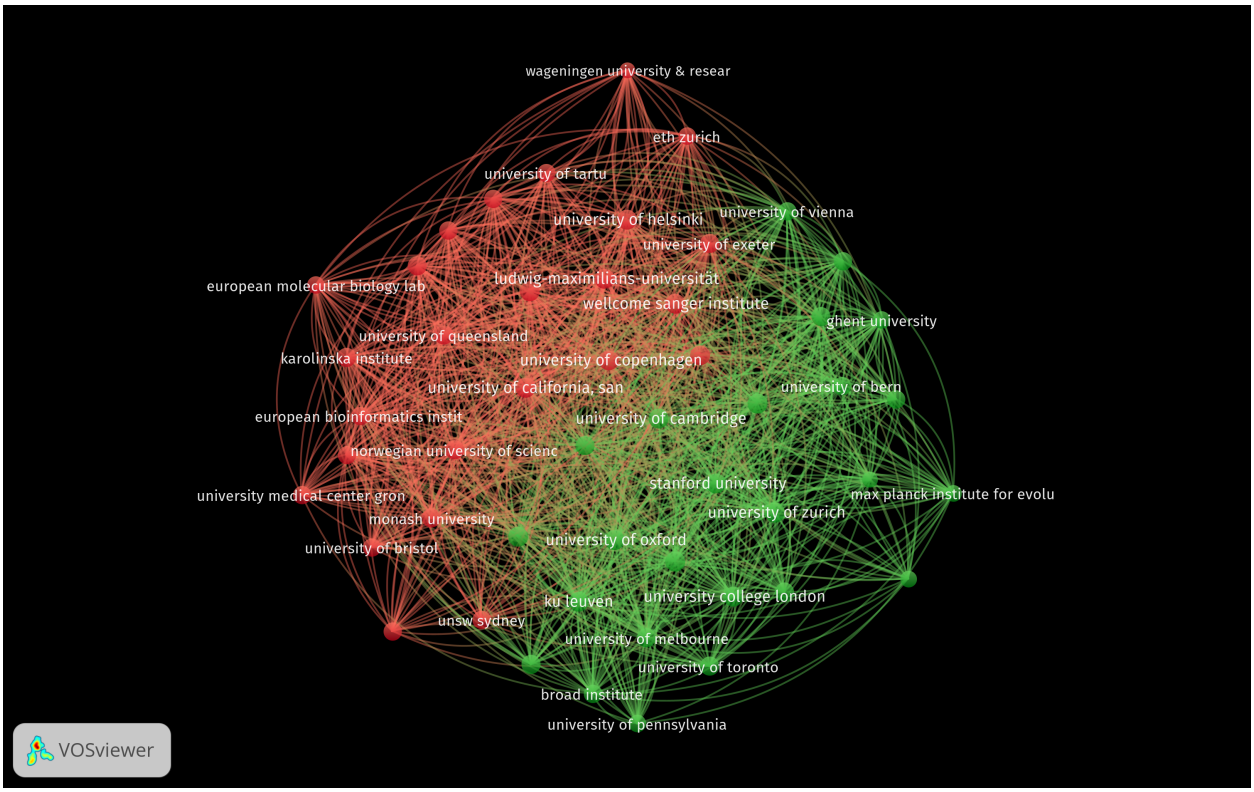

Figure 6: Data Governance - Open Science - Covid-19 - Epistemic Diversity

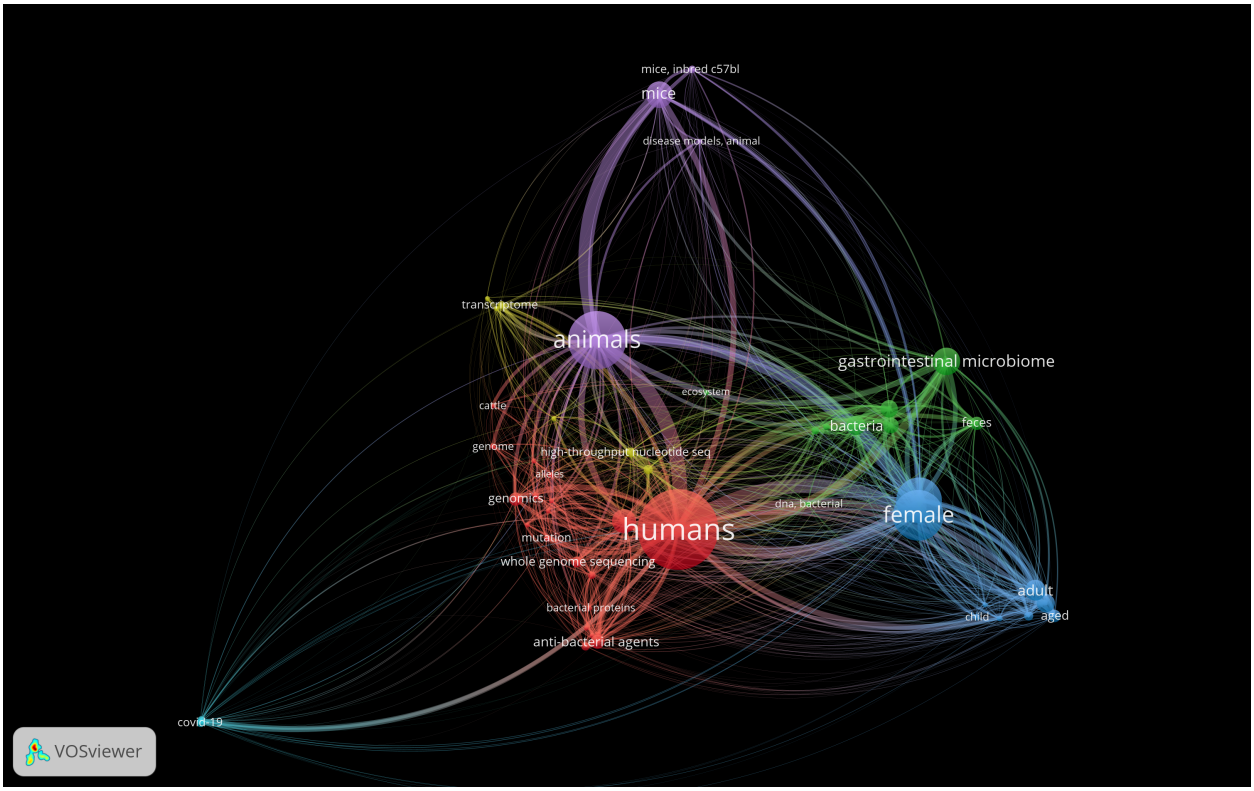

Figure 7: Data Governance - Open Science - Covid-19 - Epistemic Diversity