
R IN TRUST - A CASE STUDY OF SARS-CoV2 OPEN DATA REPOSITORIES

A PREPRINT

Nathanael Sheehan

College of Engineering, Mathematics and Physical Sciences
Exeter University
University of Exeter, St Germans Road, EX4 7HD Exeter, UK
`ns651@exeter.ac.uk`

Sabina Leonelli

EGENIS, Centre For The Study Of Life Sciences
Exeter University
University of Exeter, St Germans Road, EX4 7HD Exeter, UK
`s.leonelli@exeter.ac.uk`

March 23, 2022

Abstract

In this paper I argue on the importance of TRUST

Keywords SARS-CoV2 · Open Science · Data Governance

1 Introduction

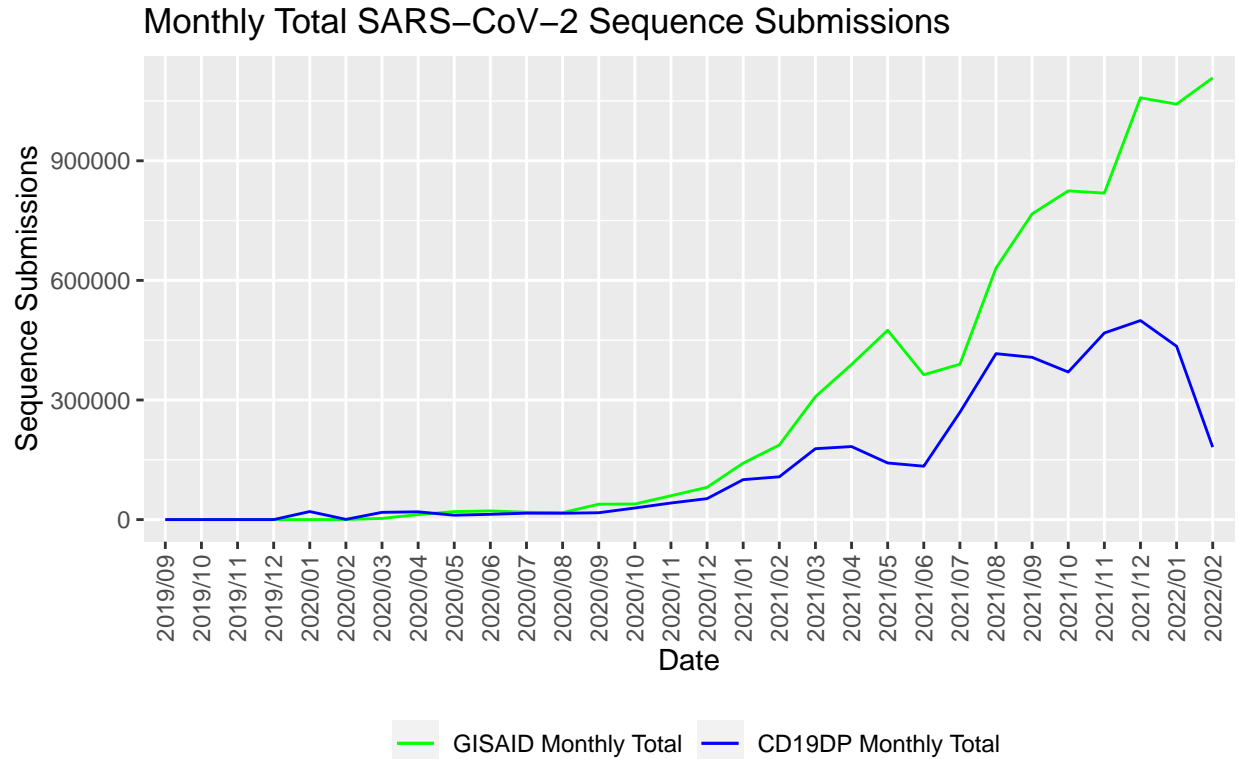
Open Science has transformed the classical modus operandi of scientific knowledge accumulation and dissemination. The pandemic highlighted the necessity for rapid, scalable and open access to the latest research findings, treatments and protocols on the coronavirus. Now exists hundreds of thousands data repositories storing data on the biological entity. Data repositories have been built by Governments, universities, businesses and publics. The greatest example of this has been the sharing of SARS genomic data, where millions of genomes have been shared across the world. During the Pandemic an open letter was written by the EBI in support in complete openness in genome sharing and asked for people to donate all data to a trust of databases. Many public figures and institutions supported the letter saying that covid was bigger than all of us, and we must come together. Others criticised the letter and argued that there database was a wolf in open source clothing and missed the point in responsibility for data sharing.

This shift in research practice - in conjunction with decreasing costs in data storage - has led to an exponential increase in “big” data and public repositories that store such data. This new found importance for data repositories means

TRUST is a set of principles for data repository owners to adhere to so the community is served and they demonstrate the ability to manage the data they hold. T = transparency, U = User focus, S = Sustainability, T = Trust, R = Adhering to the designated community’s metadata and curation standards, along with providing stewardship of the data holdings e.g. technical validation, documentation, quality control, authenticity protection, and long-term persistence.

Providing data services e.g. portal and machine interfaces, data download or server-side processing.

Managing the intellectual property rights of data producers, the protection of sensitive information resources, and the security of the system and its content. The FAIR Data Principles³ highlight the need to embrace



GISAID Metadata: <https://www.epicov.org/>
 Covid-19 Data Platform Metadata: <https://www.ebi.ac.uk/ena/portal/api/>

Figure 1: Monthly totals of global SARS-CoV-2 cases sequenced and shared on the GISAID and Covid-19 Data Platform database until February 22 2022

good practice by defining essential characteristics of data objects to ensure that data are reusable by humans and machines The Open Archival Information System (OAIS) reference model⁴ provides recommendations on setting up archives delivering long-term preservation of and access to information (in particular, digital information) and creating preservation packages. None of these encapsulate trust and temporal demand that trust takes to build and serve a community [trust table]

To start I document my experience joining each community and qualitatively narrate my experience conceptualising the R principles in each repo I then synthesise the datasets into aggregated months and explore statistical trends in each dataset I conclude by a final discourse on how

2 Methods

2.1 GISAID

In May of 2008 the Global Initiative on Sharing All Influenza Data (GISAID) was launched in tandem with the Sixty-first World Health Assembly. GISAID From its inception GISAID was built to be an alternative to the classical public domain sharing model as it took into account the beliefs of Member states by providing an accessible database designed by scientists for scientists.

2.2 COVID-19 Data Platform

3 Discussion

4 Conclusion