Between Open and Closed Data: A Metascientific Study on Global Health Data Practices

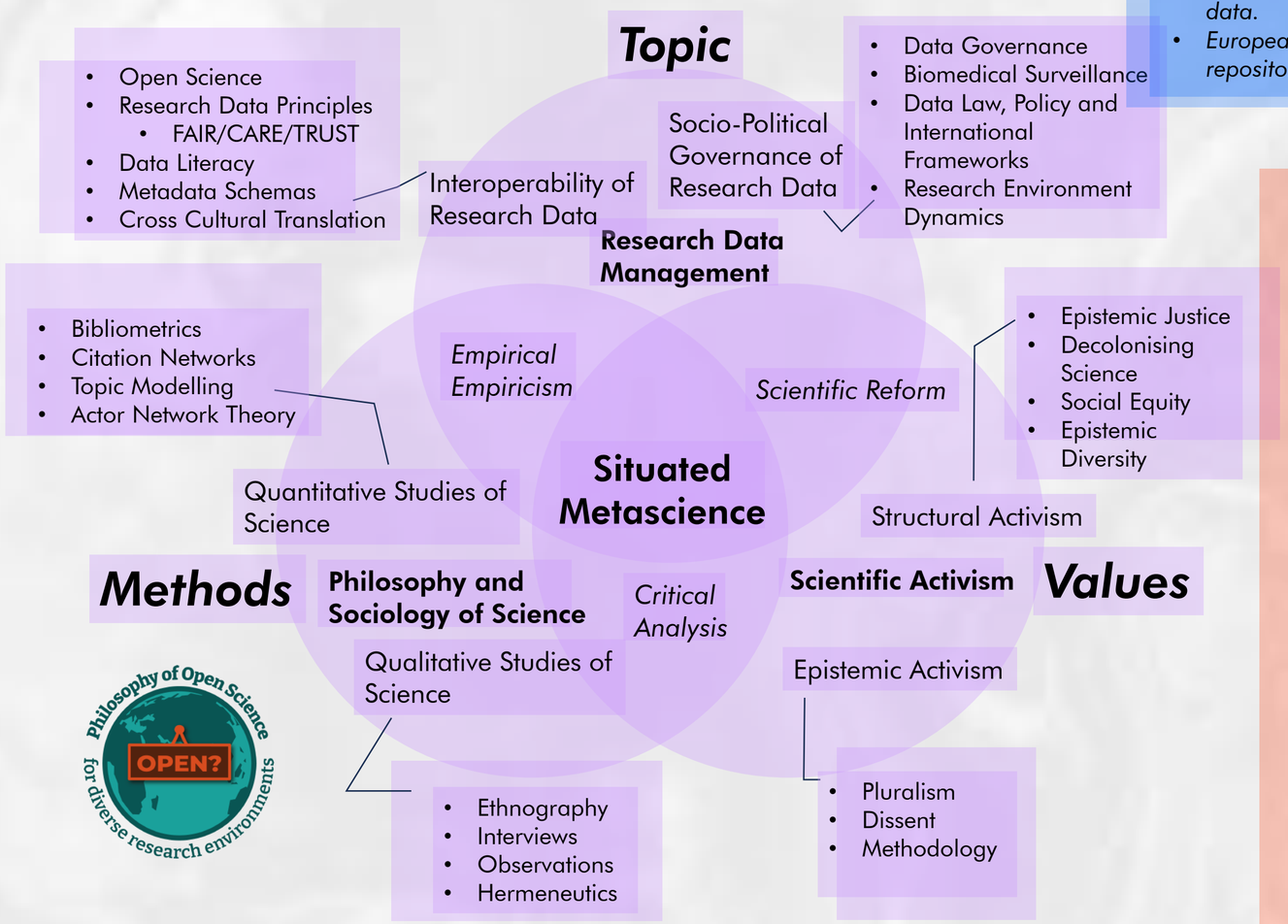
Nathanael Sheehan (ns651@exeter.ac.uk)

Research Questions

- What are the implications of openly accessible data and software for issues of access, inequity, and sovereignty in the global health landscape?
- How can situated metascience elucidate the diversity of research environments and data practices?

Research Design

Metascience is a developing field that explores three key areas: the science of science, open science, and methodological activism to enhance research practices. Unlike some metascientists who show resistance to related disciplines like the history and philosophy of science, sociology, and science, technology, and society (STS) studies, my research embraces these fields, integrating their methods and values into my research design of a "situated metascience".



Case Study (1) EMBL-EBI

Bio
EMBL-EBI is an assemblage of over forty data resources for every type of life science data. Data resources include deposition databases (which archive experimental data), added-value databases (which add value to archived data by providing annotation, curation, reanalysis and integration), as well as open-source software tools. In 2022, the open resources managed by EMBL-EBI were estimated to be worth £5.5 billion

Location: Hinxton, United Kingdom
Data Access: *Open*
Data resources include:

- *The European Nucleotide Archive – the oldest and largest DNA archive in the world.*
- *AlphaFold DB – open access database of over 200 million protein structure predictions.*
- *European Phenome Genome Archive – closed databased for human biomolecular and phenotypic data.*
- *European PubMed Central – open access literature repository*

Case Study (2) cidacs

Bio:
Established in 2016 with backing from the Brazilian Ministry of Health, CIDACS links health administrative data from more than 20 social protection programmes and government databases into high-quality, reliable longitudinal insights on epidemiological, environmental and socio-economic detriments on health. Access to CIDACS data remains restricted in efforts to build a security and ethically focused infrastructure.

Location: Salvador, Brazil
Data Access: *Closed*
Data resources include:

- *100 Million Brazilians Cohort* - one of the largest cohorts in the world focusing on low-income population).
- *Zika Platform - containing a birth cohort over a period of 30 years).*
- *CIDACS Climate Platform – a linked dataset of the two aforenoted resources with satellite imagery of the region of Bahia.*

Research Outcomes

Lesson (1)

Unrestricted | Regulated Data Access ≠ Global Representation

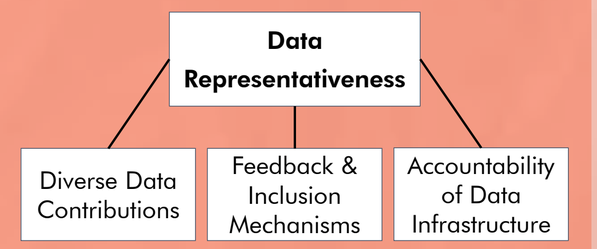
We compare the extent to which the two leading data infrastructures for SARS-CoV-2 genomic data facilitated collaboration from around the globe to understand how data reuse can enhance forms of diversity between institutions, countries, and funding groups.

Scan to read our preprint

Lesson (2)

From *Data Actionability or Data Accountability to Data Representativeness*

Key aspects of Data Representativeness
(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



Lesson (3)

Towards Decentralised Open Science:

- adoption of locally-developed software which promotes epistemic diversity and regional relevance.
- facilitation of data format conversions which enhances global interoperability and cross-cultural collaboration.
- emphasis on capacity building for financial support which nurtures local agency and sustained participation.

References: Barreto, Mauricio Lima, Laura Rodrigues, Spiros Denaxas, George Barbosa, M. Sanni Ali, Manoel Barral-Netto, Gerson Penna, et al. 2019. 'The Center for Data and Knowledge Integration for Health (CIDACS): An Experience of Linking Health and Social Data in Brazil'. International Journal of Population Data Science 4 (2). https://doi.org/10.23889/ijpds.v4i2.1140.Haraway, Donna. 1988. 'Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective'. Feminist Studies 14 (3): 575–99. https://doi.org/10.2307/3178066.Peterson, David, and Aaron Panofsky. 2023. 'Metascience as a Scientific Social Movement'. Minerva 61 (2): 147–74. https://doi.org/10.1007/s11024-023-09490-3.Sheehan, Nathanael, Federico Botta, and Sabina Leonelli. 2024. 'Unrestricted Versus Regulated Open Data Governance: A Bibliometric Comparison of SARS-CoV-2 Nucleotide Sequence Databases'. bioRxiv. https://doi.org/10.1101/2023.05.13.540634.Thakur, Matthew, Annalisa Buniello, Catherine Brooksbank, Kim T Gurwitz, Matthew Hall, Matthew Hartley, David G Hulcoop, et al. 2024. 'EMBL's European Bioinformatics Institute (EMBL-EBI) in 2023'. Nucleic Acids Research 52 (D1): D10–17. https://doi.org/10.1093/nar/gkad1088.

University of Exeter

Funding: 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 authors' views and the Commission /Agency is not responsible for any use that may be made of the information it contains. N.S. was funded via a doctoral training grant awarded as part of the UKRI AI Centre for Doctoral Training in Environmental Intelligence (UKRI grant number EP/S022074/1).

Acknowledgements: I extend heartfelt thanks to Professor Sabina Leonelli for her invaluable guidance, and the PHIL_OS team for their unwavering support and feedback on this work. Special thanks to Bianca for her continuous love and to Adrian Curry for organising the work to be publicly displayed. Their collective efforts have been crucial to our research.