The Epistemic Consequences of Open **Data and Software**: A practice based study

Nathanael Sheehan

**Supervisory team:**

Sabina Leonelli (70%)

&

Hywel Williams (30%)

**1. Objectives of** **research**:

# The main aims of this research is to explore to what extent can open data and software be used to ensure epistemic diversity, and how these tools can help in the realisation of epistemic (in)justice.

----------------------------------------------------------------------------------------------------------------------------

Research Questions:

\* How can epistemic diversity be promoted within open data and software? \*

\* What are the implications of epistemic diversity for open data and software? \*

\* How can we use open data and software to help realise epistemic (in)justice? \*

----------------------------------------------------------------------------------------------------------------------------

The aim of this thesis is to defend and explore the notions of epistemic diversity – concerning the diversification of inter alia methods, characteristics of researchers, funding, geo-political location and intellectual property regimes - and epistemic (in)justice – which refers to the idea of how knowledge is related to power, including issues surrounding oppression and various forms prejudice – as useful philosophical concepts which can guide open data and software to be more equitable and actionable. There are two main reasons why paying attention to epistemic diversity and epistemic justice are important when considering open data and software: first, if digital technologies are to benefit everyone equally then it is crucial that they take into account different types of knowledge and experience: and second, to avoid exacerbating existing inequalities then it is important to be aware of how power relations can influence the ways in which data is collected, disseminated and used. While the debate concerning epistemic diversity and epistemic justice has been studied regarding algorithmic decision making (REF), citizen science platforms and artificial intelligence systems, a lack of literature concentrates on the role of open data and software as an infrastructure that perpetuates these concepts. This thesis addresses these gaps by bridging together these two concepts in order to develop a more nuanced understanding of how open data and software might impact upon different groups within society differently. We ask to what extent open data and software can be used to ensure epistemic diversity, and how these tools can help in the realisation of epistemic justice? In order to address these questions, we take an interdisciplinary approach that combines methods and theory from philosophy of science and computer science in order to deliver a three-case study approach from genomic, ? and environmental science, with the aim to explore the hugely diverse and contested nature of knowledge in a range of open data and software practices.

# **2. Context**

The following section aims to explore the relevant body literature at the intersections of this thesis.

**2.2 (Open) Science**

**2.2.1 An Epistemic Shift**

The hegemony of Western science is being challenged by alternative knowledge traditions, and the rise of big data as a scientific power is reconfiguring global patterns of scientific collaboration and competition (Beaulieu and Leonelli 2021). Big data has radically changed how science is practised, both in terms of pedology and scientific method. Leonelli (2016) stresses that scientists must be aware of the epistemological implications of this shift. These new epistemologies are data-centric, algorithmic, and networked and have important implications for the way science is conducted (Kitchin 2014). Data-centric epistemologies focus on the collection, management, and analysis of data. Algorithmic epistemologies focus on the use of algorithms to process data and extract knowledge. Networked epistemologies focus on the use of networks to share data and knowledge. One new paradigm shift is that big data changes the way we do science by altering the relationship between data and theory. In the past, scientists would develop a theory, and then look for data to support it – now with big data, scientists can now let the data lead them to new theories (Mazzocchi 2015). Other paradigm shifts associated with big data are from data to knowledge; certainty gives way to uncertainty - as big data reveals the limits of our knowledge - and reductionism is replaced by holism - as big data finds the complex interconnections between the ever growing sea of data elements (Kitchin 2016).

These new epistemologies have increased debates surrounding the reputation of science in society. The “reproducibility crisis” in computer science (Peng 2015, Wong et al 2016, Goodman et al 2016), the “replication crisis” in the life sciences (Maxwell et al 2015), and the “crisis of confidence” in economics (Christensen and Miguel 2018) are just a few examples of the challenges threatening the credibility of science. In this context, the concept of open science has emerged as a response to the perceived need for greater transparency, inclusivity, and accountability in the conduct of science. The term “open science” has been multifariously defined by scientific practitioners (Levin et al 2016), but generally refers to a movement toward sharing all the data (Open Data), methods (Open Source), materials (Open Access) and tools (Open Software) involved in the research process, and to do so during rather than after the research is completed (Foster and Deardorff 2017). Openness in science is often contrasted with the “closed” or “secretive” practices of traditional science, which are seen as exclusive, elitist, and unaccountable (Lovelace 2021). In this sense, Open science is motivated by the belief that more open practices will lead to more rigorous, transparent, and reproducible research, and that this in turn will lead to better science and more public trust in science (Woelfle et al 2011; McKiernan et al 2016; Burgelman et al 2019).

**2.2.1 Being Open About Openness**

An emphasis on decolonality and epistemic justice come to light among many of the presentations at the UN Open Science conference 2021, which concomitantly call into reflection traditional scientific conservatism and the new imaginaries possible for epistemic justice and societal change. By reminding us that open does not equal fair, Natalia Norori confronts the euro-centric model of thinking that science should be made somehow fair and reminds us through her work on the escalating refugee crisis in Nicaragua that openness in science can contribute to equity only **if**

"[Open Science] *enables people who are historically marginalized to learn about and research topics important to them and their communities, and have their research recognized and rewarded, and have that translate in impact for their communities*"

Here, Norori reminds us that the status quo of openness reinforces existing structures of power and possibility while simultaneously solidifying epistemic injustice among marginalised knowers. De-colonial ways of questioning are what Antoinette Foster articulates as a critical component in shifting our path to epistemic justice in Open Science. Through her insights, Foster breaks down the structures in which individuals and groups have reinforced decisions rooted in harmful values and thereby accelerated class, racial and global inequalities and systemic oppression. To address these issues Babini and Rovelli research in South America calls "for more international collective action" through building a "collective international voice" to address systemic issues facing a global society. In this way, the work mentioned foregrounds how open practices can create opportunities for epistemic diversity and justice - but also underscores that these practices are not innately fair or just. Rather, they must be critically examined and continually reworked to ensure that all voices are heard and that scientific research benefits everyone, not just those in power.

While the vast majority of Open Science scholarship continues to be written and studied from a Western perspective (Ross-Hellauer et al 2022), it is important to remember that open science is not just a Western phenomenon; Open Science is a global phenomenon with important implications for the future of science and technology (Kanwar et al 2010;Czerniewicz and Goodier 2014;Serwadda et al 2018;Dutta 2020a; Dutta 2020b). Dutta et al (2021) are critical of the way that the "crisis of reproducibility" and the corresponding conversation on "questionable research practices" have framed the hegemonic conversation on open science in Communication Studies. They argue that these technologies of disciplining are undemocratic and erase possibilities of knowledge claims generated by scholars from the Global South. Ross-Hellauer et al (2022) synthesise the current body of (English) Open Science literature in order to uncover the dynamics of cumulative advantage and threats to equity. There analysis tells that the normative characteristics of Open Science – Open access, Open and FAIR data, Open methods and infrastructure and Open evaluation - are not so radical, and may in fact enable “the further neoliberalisation and commodification of research knowledge”. The lack of diversity in perspectives leads to a lack of understanding of how Open Science is practised in different parts of the world, which limits our ability to build upon existing knowledge and create new knowledge. This lack of understanding also has implications for the future of science, as it limits our ability to anticipate and address the challenges that global science will face in the coming years.

**2.3 Epistemic Diversity**

Epistemic diversity is generally agreed among social epistemologists as a means of producing new scientific theories (creativity) and challenging existing scientific paradigms (dissent). As a theory, epistemic diversity has been popularised by feminist philosophy of science discourse who dismiss the *lassier faire* approach – also known as the the invisible hand where no intervention in increasing diversity is thought necessary. The two factors of creativity and dissent bevel the possibilities of knowledge generation in a pluralistic fashion by altering the discourse surrounding accepted or normative scientific theories and practices (Solomon 2007). Firstly, dissent, challenges the hegemony of a particular scientific theory or practice by offering an altered projection of theories or strategises worth pursing. This can happen in a number of ways such as through criticism (Mill,Popper, Feyerband Longino), division of cognitive labour (Kitcher, Solomon), and distribution of knowledge (Harraway, Longino). Secondly, creativity, broadens the range of what count as legitimate scientific theories and practices by introducing new ideas, methods, and perspectives. This approach 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). Both dissent and creativity are important in epistemic diversity because they help to create a space for alternative ways of knowing and doing science. Without dissent, the dominant scientific paradigm would never be challenged and new perspectives would never be introduced. While, creativity is important as it helps to expand the range of what is considered to be science. By introducing new ideas and methods, epistemic diversity challenges the hegemony of traditional science and opens up new possibilities for knowledge generation (Ghosh 2021;Menezes de Souza 2021).

**2.4 Epistemic Injustice**

Epistemic injustice is an epistemologically significant form of social injustice which occurs when someone is deprived in their capacity to gain knowledge and justified belief because of unjust power relations. Toward epistemic justice, we must attend to the ways in which scientific practices are not neutral, but rather embedded within structures of power that produce and reproduce epistemic injustice. As a philosophical mode of inquiry, epistemic injustice has been popularised by Miranda Fricker’s (2007) notions of testimonial injustice – related to the trusting of someone’s claim as being a knower – and hermeneutical injustice – related to the how people understand their own lives. Since Frickers initial contribution, a number of philosophers have developed the idea of epistemic injustice to focus on the role of recognition or thus a lack of it (Congdon 2017;Jackson 2018;Giladi 2018).As Matthew Congdon argue, the analogy “between epistemic and moral perception should be modified to indicate a closer relationship than mere analogy” as recognition demands a number of ethical considerations (Congdon 2018). Epistemic injustice has also been associated with open access, digital environments and data science more broadly. Gloria Origgi and Serena Ciranna start this contribution by identifying how the omnipresence of ICT infrastructure and the institutions who manage them have created new forms of both testimonial injustice and hermeneutical injustice by denying people their credibility as a knowers. Knöchelmann (2021) critiques the notion that open access enables the democratisation of scientific knowledge and instead solidifies notions of epistemic injustice as it neglects the delicate intricacies of cultural and hegemony. While Symons and Alvarado (2022) identify through a case study of the criminal justice system, how methods utilised in data science enable a permanent diminution of someone ability to respect their understanding as a knower.

**2.5 Research Opportunity**

The current literature lacks a clear understanding of how open data and software can help to ensure epistemic diversity and epistemic justice. This thesis aims to address this gap by providing both an ethical framework and critical review of open science initiatives that can be used to assess the potential impact of open data and software on epistemic diversity and epistemic justice.

# **3. Activities, Case-Studies, Data, Methods and Planned Work:**

The three crises of biodiversity, climate and emerging infectious diseases are all interconnected with each other and situate themselves in the temporality of this thesis. The loss of biodiversity is a major driver of climate change, as well as being a direct cause of zoonotic diseases. Climate change then amplifies the risk of zoonotic diseases by changing the ecology of vector-borne and water-borne pathogens and making them more difficult to control. Open data and software will play a crucial role in shaping the solutions to these crises. The availability of open data will allow for greater transparency and collaboration between different communities, while the use of open source software will facilitate the development of new tools and applications to address these challenges. This research utilises a three-case study approach from these crises in order to explore the hugely diverse and contested nature of knowledge in a range of open data and software practices. The three case studies were selected because they represent different types of knowledge production, different geographical locations, and different types of open data and software initiatives. In order to keep “theory, practice and the world simultaneously, and never in isolation from each other” (Ankey et al 2011), a practice-based method for philosophy of science will be undertaken based on previous methods derived from The Society for Philosophy of Science in Practice (SPSP). Practice based philosophy of science is well suited to studying complex socio-technical systems, such as open data and software initiatives, which are often too general to study using traditional methods such as surveys or experiments.

**3.1 WP1: GISAID and The Covid-19 Data Portal**

**3.1.1 Introduction**

The rapid dissemination of genomic data on the SARS-CoV-2 virus has been crucial for understanding the viruses evolution and for developing effective vaccines, protocols treatments. However, the way in which this data is shared has been far from uniform, with different databases adopting different approaches to responsible openness. On the 29th of January 2021 the governing board of the European Bioinformatics Institute (EBI) posted a public letter calling for a greater “openness” in sharing SARS-CoV-2 genome data (EMBL 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 (EBI, The GenBank of USA and the DNA Data Bank of Japan). The letter was signed and promoted by Nobel Laureates, Directors of Bioinformatic programs and many researchers at the cutting edge of genome sequencing. At the same time, the Global Initiative on Sharing Avian Influenza Data (GISAID) had just overtaken the EBI’s European COVID-19 Data Portal (C19DP) in the volume of genome sequences being shared to open access databases. GISAID was launched in 2008 to monitor global influenza outbreaks and from the offset positioned itself as an alternative to the public domain sharing model. Its policy requires users to authenticate their identity and agree not to republish or link GISAID genomes with other datasets without permission from the data producer (GISAID 2022). This requirement stems from the recognition that some researchers – often working in low-resourced environments and/or less visible research locations – are reluctant to share data due to fears of better-equipped researchers building on such work without due acknowledgment (Emanuel 2004; Bezuidenhout and Chakauya 2018). Indeed, the GISAID model has fostered trust and information exchange among groups that differ considerably in their geo-political locations, funding levels, material resources and social characteristics, thereby expanding the range of data sources shared online. This proved decisive when, at the beginning of 2020, GISAID launched the EpiCov database which stores, analyses and builds evolutionary trees of SARS COV-2 genome sequences – now the leading open access database for SARS-CoV-2, with over 9 million genomes sequenced by April 2022. At the same time, by limiting the extent to which data can be accessed and linked, GISAID has been criticised for negatively affecting the insight, pace and breadth of future research on SARS-Cov2 – leading to the backlash by hundreds of leading researchers concerned about the urgency of an effective pandemic response (Gozashti and Corbett-Detig 2021;Wadman 2021; see: Collins Thread 2021;Van Noorden 2021;Yehudi 2022).

**3.1.2 Planned Work and Methods**

--------------------------

**Paper 1:**

LIT: A responsible open data framework

--------------------------

In recent years, there has been an increased focus on the responsible sharing of data (Carroll et al 2020; Stoyanovich 2020; Gebru et al 2021;). This is in part due to the recognition of the potential benefits of data sharing for science and society, as well as the need to address some of the challenges associated with data sharing such as privacy and intellectual property (Leonelli 2021;Samlali and Stern 2021). By far the most popular, the FAIR principles have been widely endorsed by funders and institutions, and are increasingly being used as a framework for thinking about how to improve the accessibility and re-usability of research data (Wilkinson et al., 2016). The framework has grown rapidly in recognition in and outside of academia, with over 7100 publications citing it as of July 2021 and now includes FAIR research software (Lamprecht et al 2020, Katz et al 2021). While the FAIR principles are often presented as a set of guidelines for researchers, they also have implications for data governance more broadly (REF). In particular, the FAIR principles could be seen as a way of ensuring that data is responsibly managed and used (Bezuidenhout 2020). However, we argue that the FAIR principles are not focused on responsibility at all. To make this point, we discuss that the conceptualisation and implementation of the FAIR principles reflects a narrow construed idea of what responsibility might be for data and instead the FAIR principles focus on data usability rather than a response to the broader context of data sharing and data reuse. There is a risk that the narrow focus on usability could lead to a situation where responsibility for data is seen as primarily an issue for individual researchers or institutions, rather than something that needs to be addressed at a structural level (REF). In order to understand why the FAIR principles are not focused on responsibility, it is first necessary to define what we mean by “responsibility”. We use the following definition of responsibility taken from Lin et al (2020) work on the TRUST principles:

“*repositories take responsibility for the stewardship of their data holdings and for serving their user community. Responsibility is demonstrated by:*

* *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 mean 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.”*

Building on this notion of responsibility we propose a conceptual framework consisting of three In this paper, we propose a conceptual framework for responsible data sharing that builds on existing work in the area (Wilkinson 2016;Lin et al 2021;). The framework consists of three key elements: legalities, interpretability and tractability. Each element is associated with a set of assessment criteria that can influence epistemic diversity and epistemic justice. The assessments for each element capture: legalities (concerning data sharing, including dealing with intellectual property (IP) and managing consent and privacy), tractability (accessing, handling, working with and manipulating data objects) and interpretability (data provenance, allowing the ability to interpret data, to contextualise and find meaning from results). For each element is assigned a test where a score is given based on the criteria. Each metric varies on a scale from 0.9, where 0 indicates the responsibility classification is “not implemented”, 0.4 indicates the responsibility classification is “partially implemented” and 0.9 indicates the responsibility classification is “sufficiently implemented”. The total score for each distinct element is calculated as a mean weighted average from each of its sub classifications. The assessments are designed to be a starting point for thinking about how research practices can contribute to more responsible and diverse research practices. We acknowledge that the framework is not exhaustive of all important issues that pertain to responsibility in data science, but rather, offer it as a useful tool for thinking through some of these often neglected areas.

--------------------------

**Paper 2:**

Diverse Enough? An empirical investigation into epistemic diversity within genomic surveillance at the institutional level

--------------------------

The SARS-COV2 pandemic has resulted in a massive increase in the sharing of genomic data. However, this data sharing has not been without its problems. In particular, there have been concerns about the way in which data has been shared and used. In this paper I will argue that the way in which SARS-COV2 genome data has been shared and used reflects a lack of understanding of the ethical implications of such sharing, as well as a failure to consider the unequal power relations between different actors involved in the scientific process. To make this point I will draw on a range of data-centric methodologies, from network visualisation of key actors involved in data sharing policies, network and textual analysis from a large corpus of genomic surveillance papers and a Bayesian nonparametric regression model of institutional diversity indicators.

**Network Visualisation**

I will firstly present a network visualisation of key actors involved in SARS-COV2 genome data sharing based on the Collins Thread – an openly shared email chain between GISAID and the Heads of International Research Organizations. This analysis reveals a number of important patterns, including the fact that there is a small number of highly centralised actors who are disproportionately influential in setting policy for data sharing. These actors are predominantly from high-income countries, and their influence is likely due to their greater financial resources and scientific expertise. This analysis also reveals that there are very few connections between these highly centralised actors and those from low- and middle-income countries, which suggests that there is a lack of communication and collaboration between these groups.

**Network and textual** **Analysis**: In the next section of this paper I will use methods from network science and natural language processing to analyse relationships between different groups of researchers involved in genomic surveillance. To do this I will collect full texts and metadata using the web of science API containing the term “genome | genomic surveillance” from the post genomic era to present. The data will be formatted and cleaned and subsequently analysed by curating summary statistics as well as generating three bibliographic bipartite networks. Within the summary statistics, annual scientific production, annual percentage growth rate, most productive authors, top citations, corresponding authors counties, total citations per county, most relevant sources, and most relevant keywords will all be calculated. The three bipartite networks will be structured based on county scientific collaboration, keyword co-occurrence and co-authorship frequency. In order to provide a descriptive analysis of network characteristics several statistical measurements will be taken. Firstly, in order to describe the structural properties of a network a set of summary statistics will be calculated (see table 1 summary statistics). Secondly, additional measurements will be calculated to identify the most important vertices in a given network, as well other vertex level measurements (see table 1 vertices). Finally, in order to identify the most important subgroups within a network I will explore a range of community detection algorithms. These are methods that divide a network into groups of vertices according to some similarity measure between the vertices within each group. The aim is to partition a network into communities such that the number of connections between nodes within communities is maximally high, while connections between nodes in different communities are minimally high (see table 1 community detection).

Table 1: Network Statistics

|  |  |  |
| --- | --- | --- |
| **Summary Statistics** | **Vertices** | **Community Detection** |
| Size | Degree centrality | Modularity maximisation |
| Density | Closeness centrality | Edge betweenness optimisation |
| Transitivity | Eigenvector centrality | Label propagation algorithm |
| Diameter | Betweenness centrality | Info-map algorithm |
| Degree Distribution | PageRank score | Leading eigenvector method |
| Degree centralisation | Hub score | Spin glasses method |
| Closeness centralisation | Authority score |  |
| Eigenvector centralisation | Vertex ranking |  |
| Betweenness centralisation |  |  |
| Average path length |  |  |

**Bayesian nonparametric regression**: There have been a number of attempts in the literature to draw correlations between the number of genomes being sequenced in a given country and some predictor value. For instance, Romano (2020) use confirmed cases per country in a simple regression to stress the importance of genomic surveillance infrastructure for dealing with the pandemic, while Faria et al (2021) considers a number of socio-economic covariates such as expenditure on R&D, GDP, a socio-demographic index and percentage of influenza cases being sequenced as a proxy to show global disparities in genomic surveillance. At the same time, Plessis et al (2021) demonstrate how epidemiological and travel strongly influence the dynamics of genome sharing. While all of these represent a plurality of way to think about what effects the action of sharing genome sequences at various geographical scales, levels of statistical rigour and embedded values – there is a lack of systematic studies on the institutional level for both disparities and diversity in the number of genomes being sequenced and shared. One possible reasons for this gap may be the multiplicity in factors that influence the number of genomes being sequenced and shared. This paper answers this uncertainty through the use of a Bayesian nonparametric regression model to examine the institutional correlates of diversity and disparity in the number of genomes being sequenced and shared across countries. Bayesian nonparametric regression models have proved useful for a number of health-care, social science, and ecological applications. The model described in this paper can be viewed as a finite mixtures of regression models that are estimated with Bayesian methods to examine the variability in genome sharing in countries with different types of institutions.

--------------------------

**Internship:**

GISAID

--------------------------

As an intern at GISAID, I would have the opportunity to collect empirical data based on ethnographic observations and interviews. I would be able to work with a diverse range of contributors and learn about their experiences working within the organisation. The internship would last 1-3 months, and at the end of it, I would produce an auto-ethnography detailing my experience.

**3.1.2 WP2: Advancing Capacity for Climate and Environment Social Science (ACCESS)**

The mission of the ACCESS project is to champion and coordinate environmental social science knowledge and data through timely and accessible communication in varied formats tailored to different stakeholder communities, integrated with partner activities The rationality behind the project is for environmental social science to maintain relevance and impact, knowledge, insight and data must be accessible to stakeholders and society. To achieve this the project will deliver new resources, tools and techniques to enhance visibility, heighten awareness and sustain the interest and use of climate and environment social science in diverse social groups. The scope, methods and inclusion for this project are still yet to be confirmed.

**3.1.3 WP3: (?)**

# **4. Ethics & Responsible Research:**

This PhD will examine the role of epistemic diversity and epistemic justice in open data and software initiatives as a way to make science more inclusive and actionable. As such, the following cultural, political and ethical concerns have been identified and discussed.

**4.1 Cultural Issues**

Open science is often seen as a way of promoting transparency and accountability, as well as increasing the availability of data and software for research purposes. However, there is also a concern that open science initiatives can perpetuate ‘data colonialism’, whereby developed countries extract data from developing countries without providing any benefit in return (Couldry 2018). There is also a worry open open science initiatives can lead to a ‘commodification’ of scientific knowledge, whereby the data produced by a researcher is valued as a commodity to be bought and sold, rather than as a public good (Radder 2010). Software is often seen as a way of increasing the accessibility of data, as well as facilitating the analysis and interpretation of data. However, there is also a concern that open source software can lead to a ‘digital divide’, whereby people who cannot afford to buy proprietary software are excluded from using it (James 2003). In order to mitigate these effects, this work supports a ‘capabilities approach’ in which the challenges mentioned above are situated and reframed in a more holistic and critical approach (Bezuidenhout  et al. 2017).

**4.3 Political Issues**

Open Science is often recognised as ‘a political project to an even greater extent than it is a technological one’ (Prainsack and Leonelli 2018). For example, the opening of science may be associated with democracy and equality, whereby marginalised voices are given a platform to be heard, but also whereby powerful actors also gain further power (Taylor 2014). Another example can be understood in terms of the proliferation of digital tools which create produce data; this notion opens up questions regarding who owns them, who can access them and what may be done with them (REF). Considering these issues, a great deal of reflexivity by the researcher to situate their work towards the amelioration of marginalised groups and away from trends to reshape scientific practices with neo-liberal ideals of market competition, training the work force and developing new products and services (Ross-Hellauer et al 2022).

**4.4 Ethical issues:**

The authors are all based in England, meaning that they are bound by the ethical regulations of the country. In order to adhere to these regulations, the thesis will follow a few key steps. Firstly, the research planned will obtain permission from their local ethics committee in order to conduct the research with human participants. Secondly, all participants will give their informed consent prior to taking part in any research activities. Lastly, additional measures to protect the anonymity and confidentiality of all participants by ensuring that all data is collected and stored securely.

**Research Impact**

The aim of this PhD is to increase the inclusion of under-represented groups in open data and software initiatives, as well as to make science more actionable. The outputs produced from this research will have relevant impact on policy, practice and theory.

**5.1 Policy Impact**

As open science is becoming ever more integrated into policy decision making (e.g. United Nations Open Science Policy, EU, UK GOV) the findings of this PhD could be used to inform policymaking around open data and software initiatives. It is hoped, that the work conducted throughout this PhD will help demonstrate the need for an improved design and implementation of these initiatives so that they are more inclusive and actionable, while providing useful and actionable feedback to cement this into policy.

**5.2 Practice Impact**

The findings of this PhD will also be relevant to practitioners working on or interested in open data and software initiatives. It is hoped that the work conducted throughout this PhD will help to improve the design and implementation of these initiatives so that they are more inclusive and actionable. The feedback provided by this research could also help to improve the way in which these initiatives are communicated and marketed, so as to better engage potential participants from under-represented groups.

**5.3 Theory Impact**

This PhD will contribute to theory around Open Data, Software Studies, Science & Technology Studies (STS) and Inclusion/Diversity by providing empirical evidence of how under-represented groups interact with, make use of, and benefit from open data and software initiatives. The work conducted throughout this PhD has the potential to challenge current assumptions about who uses or needs open data and software, as well as what counts as valuable contributions to these fields.

# **Time Management plan:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Year 1** | | | | **Year 2** | | | | **Year 3** | | | | Submisson |
| Activities | Progress | Deadline | Sep-22 | Dec-22 | Mar-23 | Jun-23 | Sep-23 | Dec-23 | Mar-24 | Jun-24 | Sep-24 | Dec-24 | Mar-25 | Jun-25 | Sep-25 |
| **Project 1: GISAID** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| *Paper 1* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Framework creation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Framework testing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Write up |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| *Paper 2* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data Analysis |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Analysis refinements |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Write up |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| *Internship* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Project 2: ACCESS** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Particpation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Write up |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Project 3: ?** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Thesis** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Main Write up |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Review |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Finalse Thesis |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Bibliography

[Bezuidenhout, L.M., Leonelli, S., Kelly, A.H., Rappert, B., 2017. Beyond the digital divide: Towards a situated approach to open data. Science and Public Policy 44, 464–475. https://doi.org/10.1093/scipol/scw036](https://doi.org/10.1093/scipol/scw036)

Big Data, new epistemologies and paradigm shifts - Rob Kitchin, 2014 [WWW Document], n.d. URL <https://journals.sagepub.com/doi/10.1177/2053951714528481> (accessed 5.16.22).

Burgelman, J.-C., Pascu, C., Szkuta, K., Von Schomberg, R., Karalopoulos, A., Repanas, K., Schouppe, M., 2019. Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century. Frontiers in Big Data 2.

Christensen, G., Miguel, E., 2018. Transparency, Reproducibility, and the Credibility of Economics Research. Journal of Economic Literature 56, 920–980. <https://doi.org/10.1257/jel.20171350>

Congdon, M., 2018. “Knower” as an Ethical Concept: From Epistemic Agency to Mutual Recognition. Feminist Philosophy Quarterly 4. <https://doi.org/10.5206/fpq/2018.4.6228>

Czerniewicz, L., Goodier, S., 2014. Open access in South Africa: A case study and reflections. South African Journal of Science 110, 01–09. <https://doi.org/10.1590/sajs.2014/20140111>

Data and Society | SAGE Publications Inc [WWW Document], n.d. URL <https://us.sagepub.com/en-us/nam/data-and-society/book269709> (accessed 9.6.22).

Devezer, B., Nardin, L.G., Baumgaertner, B., Buzbas, E.O., 2019. Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. PLOS ONE 14, e0216125. <https://doi.org/10.1371/journal.pone.0216125>

Dutta, M., 2020. Culture-Centered Approach to Communication for Social Change, in: Dutta, M. (Ed.), Communication, Culture and Social Change: Meaning, Co-Option and Resistance, Palgrave Studies in Communication for Social Change. Springer International Publishing, Cham, pp. 239–281. <https://doi.org/10.1007/978-3-030-26470-3_6>

Dutta, M., Ramasubramanian, S., Barrett, M., Elers, C., Sarwatay, D., Raghunath, P., Kaur, S., Dutta, D., Jayan, P., Rahman, M., Tallam, E., Roy, S., Falnikar, A., Johnson, G.M., Mandal, I., Dutta, U., Basnyat, I., Soriano, C., Pavarala, V., Sreekumar, T.T., Ganesh, S., Pandi, A.R., Zapata, D., 2021. Decolonizing Open Science: Southern Interventions. Journal of Communication 71, 803–826. <https://doi.org/10.1093/joc/jqab027>

Dutta, M.J., 2020. Whiteness, internationalization, and erasure: decolonizing futures from the Global South. Communication and Critical/Cultural Studies 17, 228–235. <https://doi.org/10.1080/14791420.2020.1770825>

Fehr, C., 2018. Feminist Philosophy of Biology, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

Foster, E.D., Deardorff, A., 2017. Open Science Framework (OSF). J Med Libr Assoc 105, 203–206. <https://doi.org/10.5195/jmla.2017.88>

Fricker, M., 2007. Epistemic Injustice: Power and the Ethics of Knowing. Oxford University Press, Oxford. <https://doi.org/10.1093/acprof:oso/9780198237907.001.0001>

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K., 2021. Datasheets for Datasets. <https://doi.org/10.48550/arXiv.1803.09010>

Ghosh, B., Ramos-Mejía, M., Machado, R.C., Yuana, S.L., Schiller, K., 2021. Decolonising transitions in the Global South: Towards more epistemic diversity in transitions research. Environmental Innovation and Societal Transitions, Celebrating a decade of EIST: What’s next for transition studies? 41, 106–109. <https://doi.org/10.1016/j.eist.2021.10.029>

Goodman, S.N., Fanelli, D., Ioannidis, J.P.A., 2016. What does research reproducibility mean? Sci Transl Med 8, 341ps12. <https://doi.org/10.1126/scitranslmed.aaf5027>

Grasswick, H., 2018. Feminist Social Epistemology, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

How Do Scientists Define Openness? Exploring the Relationship Between Open Science Policies and Research Practice - Nadine Levin, Sabina Leonelli, Dagmara Weckowska, David Castle, John Dupré, 2016 [WWW Document], n.d. URL <https://journals.sagepub.com/doi/full/10.1177/0270467616668760> (accessed 6.8.22).

James, J., 2003. Free software and the digital divide: opportunities and constraints for developing countries. Journal of Information Science 29, 25–33. <https://doi.org/10.1177/016555150302900103>

Kanwar, A., Kodhandaraman, B., Umar, A., 2010. Toward Sustainable Open Education Resources: A Perspective From the Global South. American Journal of Distance Education 24, 65–80. <https://doi.org/10.1080/08923641003696588>

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

Leonelli, S., 2021. Open Science and Epistemic Diversity: Friends or Foes?

Leonelli, S., 2016. Data-Centric Biology: A Philosophical Study, Data-Centric Biology. University of Chicago Press. <https://doi.org/10.7208/9780226416502>

Leonelli, S., Lovell, R., Wheeler, B.W., Fleming, L., Williams, H., 2021. From FAIR data to fair data use: Methodological data fairness in health-related social media research. Big Data & Society 8, 20539517211010310. <https://doi.org/10.1177/20539517211010310>

Longino, H., 2019. The Social Dimensions of Scientific Knowledge, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

Longino, H., n.d. Science As Social Knowledge | Department of Philosophy [WWW Document]. URL <https://philosophy.stanford.edu/publications/science-social-knowledge> (accessed 8.4.22).

Macleod, C., 2020. John Stuart Mill, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

Maxwell, S.E., Lau, M.Y., Howard, G.S., 2015. Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? American Psychologist 70, 487–498. <https://doi.org/10.1037/a0039400>

Mazzocchi, F., 2015. Could Big Data be the end of theory in science? EMBO Rep 16, 1250–1255. <https://doi.org/10.15252/embr.201541001>

McKiernan, E.C., Bourne, P.E., Brown, C.T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B.A., Ram, K., Soderberg, C.K., Spies, J.R., Thaney, K., Updegrove, A., Woo, K.H., Yarkoni, T., n.d. How open science helps researchers succeed. eLife 5, e16800. <https://doi.org/10.7554/eLife.16800>

Menezes de Souza, L., 2014. Epistemic Diversity, Lazy Reason and Ethical Translation in Post-Colonial Contexts: the case of indigenous educational policy in Brazil. Interfaces Brasil/Canadá 14, 36–60.

Norori, N., n.d. My journey through an inequitable system 20.

Origgi, G., Ciranna, S., 2017. Epistemic Injustice: The case of digital environments, in: The Routledge Handbook of Epistemic Injustice. Routledge.

Peng, R., 2015. The reproducibility crisis in science: A statistical counterattack. Significance 12, 30–32. <https://doi.org/10.1111/j.1740-9713.2015.00827.x>

Preston, J., 2020. Paul Feyerabend, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

Project MUSE - The Commodification of Academic Research [WWW Document], n.d. URL <https://muse.jhu.edu/chapter/28146/pdf> (accessed 9.6.22).

Ross-Hellauer, T., Reichmann, S., Cole, N.L., Fessl, A., Klebel, T., Pontika, N., n.d. Dynamics of cumulative advantage and threats to equity in open science: a scoping review. Royal Society Open Science 9, 211032. <https://doi.org/10.1098/rsos.211032>

Samlali, K., Stern, J., n.d. Towards Responsible Genomic Surveillance: 27.

Science as Social Knowledge, 1990.

Serwadda, D., Ndebele, P., Grabowski, M.K., Bajunirwe, F., Wanyenze, R.K., 2018. Open data sharing and the Global South—Who benefits? Science 359, 642–643. <https://doi.org/10.1126/science.aap8395>

Stoyanovich, J., Howe, B., Jagadish, H.V., 2020. Responsible data management. Proc. VLDB Endow. 13, 3474–3488. <https://doi.org/10.14778/3415478.3415570>

Thornton, S., 2021. Karl Popper, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

Woelfle, M., Olliaro, P., Todd, M.H., 2011. Open science is a research accelerator. Nature Chem 3, 745–748. <https://doi.org/10.1038/nchem.1149>

Wong, W.E., Gao, R., Li, Y., Abreu, R., Wotawa, F., 2016. A Survey on Software Fault Localization. IEEE Transactions on Software Engineering 42, 707–740. <https://doi.org/10.1109/TSE.2016.2521368>