COVID-19: An exploration of consecutive systemic barriers to pathogen-related data sharing during a pandemic

Data for Policy 2021 paper #51

Yo Yehudi

Department of Computer Science, University of Manchester

@yoyehudi

bit.ly/data4policy-yy-covid

💌 y.yehudi@postgrad.manchester.ac.uk

COVID-19 infection and death tolls worldwide

March 2021 (when abstract authored)

> 122 million **confirmed cases**

> 2.7 million **deaths**

August 2021 (when talk prepared)

> 206 million **confirmed cases**

> 4.35 million **deaths**

Source: Our World in Data https://ourworldindata.org/explorers/coronavirus-data-explorer

Data can be surprisingly hard to **obtain**, hard to **use**, and hard to **re-distribute**.

Consecutive barriers results in poorer analyses, slower predictions, mistakes, and may make existing analyses impossible to re-share.

Methods and data used

Semi-structured interviews

Zoom-based interview, covering:

The Bad: access problems, actions when barriers encountered.

Good data sources: what did they do right, was there anything they could do better, and what would a “dream” data source look like?

Ethics: Any ethical concerns about sharing this data?

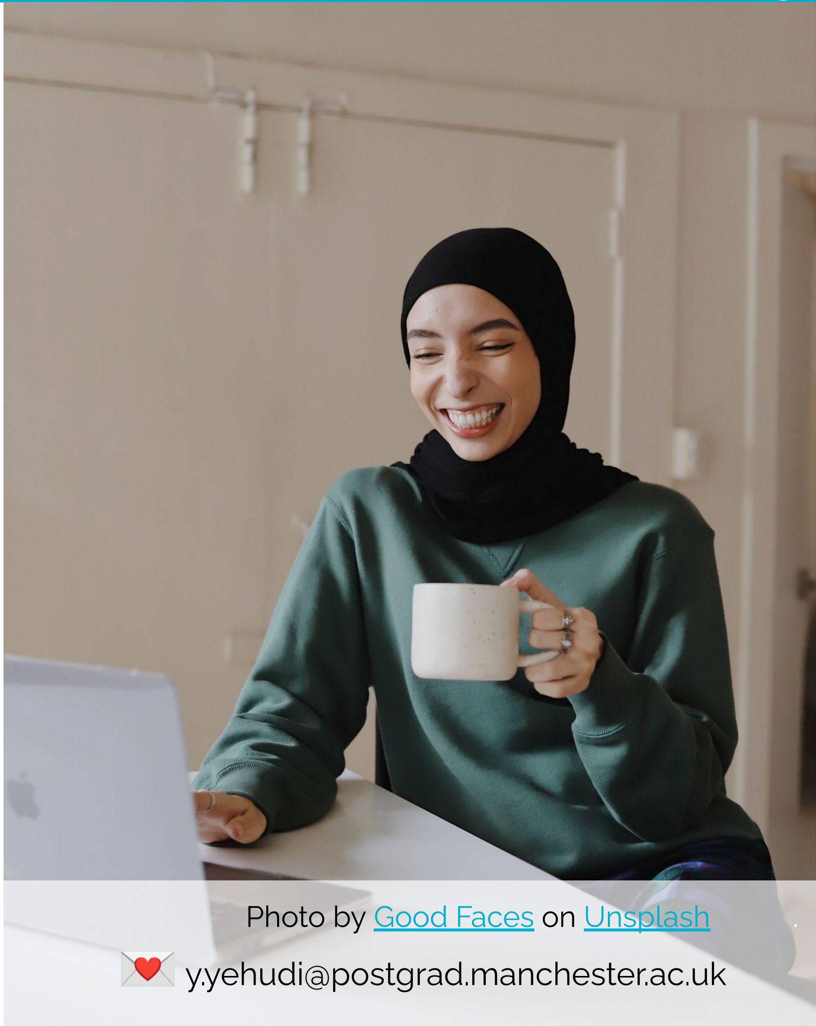


Photo by Good Faces on Unsplash

Recruitment & sampling

Complete

Round 1:

8 interviews

- Goal: speak to people working with **non-private** data (example: viral genomics yes, electronic health records, no).
- 8 participants recruited via twitter and targeted mailing list posts
- Mix of researchers, bioinformaticians, computational epidemic modellers, biologists, and healthcare professionals.
- Problem: **all** from HICs in North America & Europe.

Recruitment & sampling

Ongoing

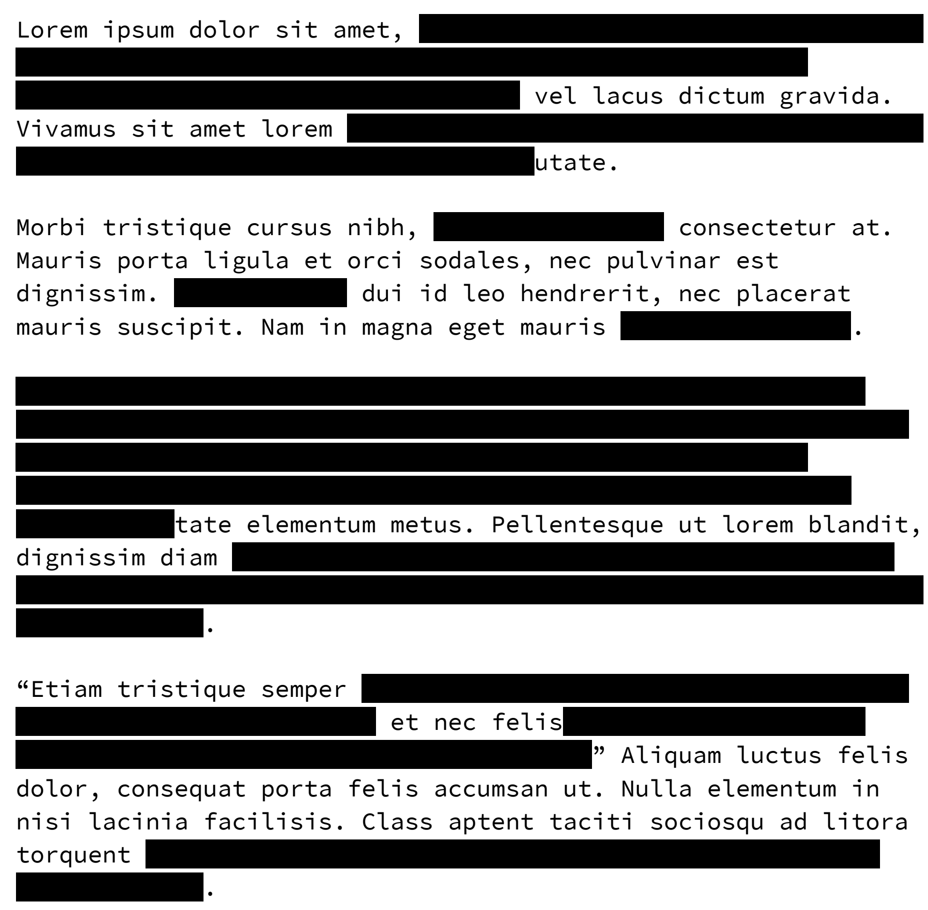
Round 2:

6+ interviews

- Now includes participants from **Asia, Africa, South America, Australasia**. Sample is now 1/3 LMIC.
- Expanded to include:
  - Civic / government data specialists
  - Open journalists
  - Software engineers building national covid reporting pipelines
  - Domestic abuse and social inequity in lockdown and pandemic conditions

Sensitive topics

Participants were given the option to review, amend, and redact their interviews.



Lorem ipsum dolor sit amet, [redacted] vel lacus dictum gravida. Vivamus sit amet lorem [redacted]utate.

Morbi tristique cursus nibh, [redacted] consectetur at. Mauris porta ligula et orci sodales, nec pulvinar est dignissim. [redacted] dui id leo hendrerit, nec placerat mauris suscipit. Nam in magna eget mauris [redacted].

[redacted]tate elementum metus. Pellentesque ut lorem blandit, dignissim diam [redacted].

“Etiam tristique semper [redacted] et nec felis[redacted]” Aliquam luctus felis dolor, consequat porta felis accumsan ut. Nulla elementum in nisi lacinia facilisis. Class aptent taciti sociosqu ad litora torquent [redacted].

Preliminary results

Based on interim analysis of Round-1 participants

Data Types - private, non-private, in between

COVID-19 data sharing barriers are found in many different domains. Many participants worked with private data as well as non-private data, and some worked exclusively with private data types.

Non-private data

YES: Data which won't expose a person's location, wellbeing, behaviours, or other personal sensitive details

YES: Data that have been shared openly with consent, that otherwise might be considered private (e.g. GPS tracks, social media posts).

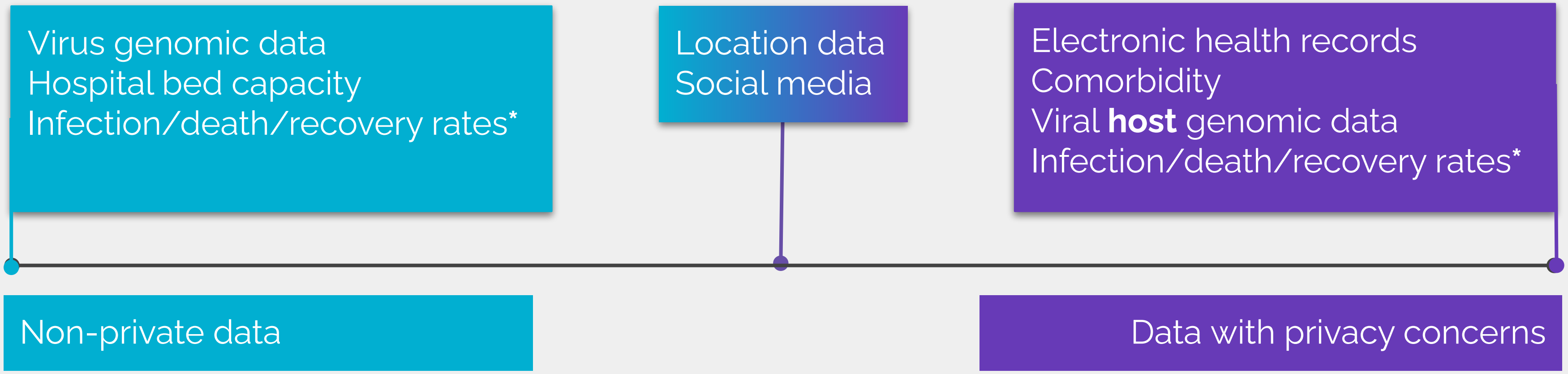
Data with privacy concerns

YES: Data which, if public, might expose a specific person's location, behaviours, wellbeing, or other personal sensitive details

NO: We're not referring to data with the potential for commercial exploitation as private.

Data Types - private, non-private, in between

COVID-19 data sharing barriers are found in many different domains. Many participants worked with private data as well as non-private data, and some worked exclusively with private data types.



Virus genomic data
Hospital bed capacity
Infection/death/recovery rates*

Location data
Social media

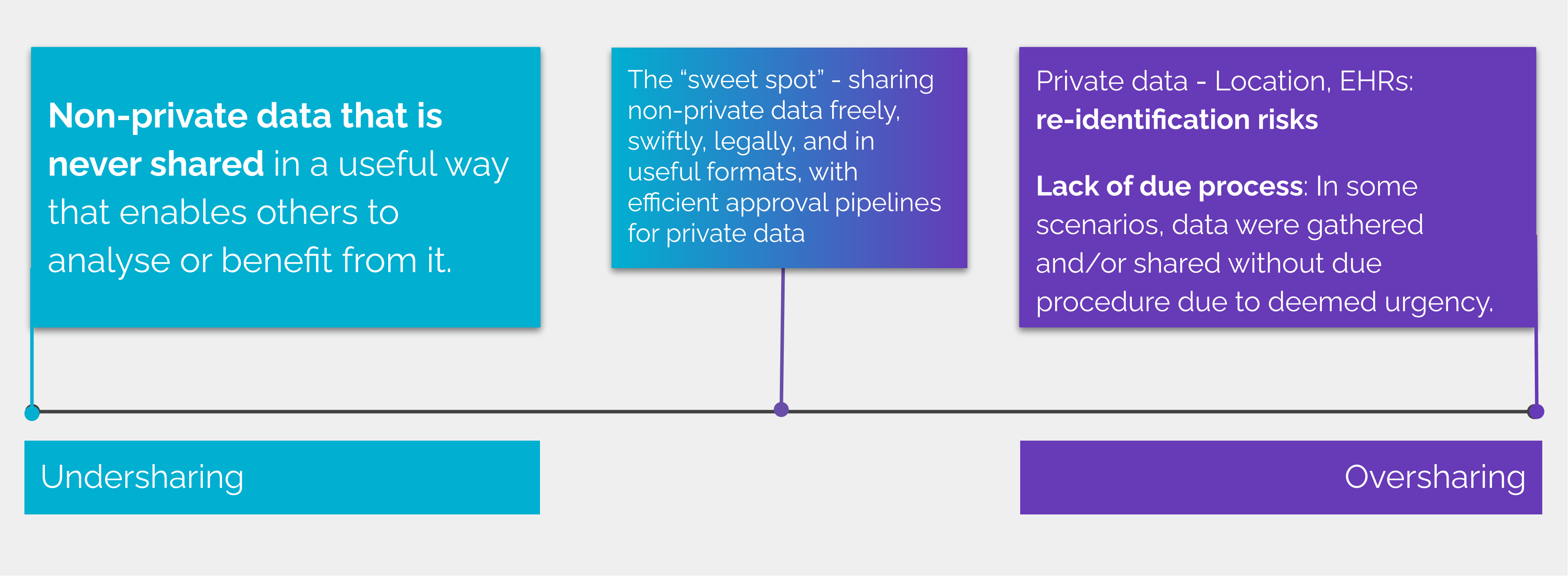
Electronic health records
Comorbidity
Viral **host** genomic data
Infection/death/recovery rates*

Non-private data

Data with privacy concerns

*may be re-identifiable in smaller populations and/or with linked datasets

Ethical concerns - a spectrum



Non-private data that is never shared in a useful way that enables others to analyse or benefit from it.

The “sweet spot” - sharing non-private data freely, swiftly, legally, and in useful formats, with efficient approval pipelines for private data

Private data - Location, EHRs: **re-identification risks**

Lack of due process: In some scenarios, data were gathered and/or shared without due procedure due to deemed urgency.

Undersharing

Oversharing

Barriers in COVID-related data sharing

3

Key barrier stages

1. Initial data **access**
2. Difficulty **using** the data once access has been granted
3. Prohibition of **re-sharing** the data and/or resultant analyses.

Barrier 1:

Data is **hard to access**

1. Data **Access** Barriers

1. **Knowing data exists** to be analysed
2. **Gatekeeping** - non-private data that is not shared - defaulting to access control
3. **Human throughput bottlenecks** in non-automatic data access pipelines
4. **Money** - both at- or below-cost data access charges and more expensive commercial charges.

> “Essentially, you had a treasure trove of information that was not at all mapped to each other, that **a lot could have been done with**, which was being **heavily access managed** and not at all curated.”

“If the mechanism for accessing data is “**email someone and wait for them to respond**”, then if you don’t ideally have both professor and OBE involved in your name, you’re going to struggle.”

Barrier 2:

Data is **hard to use**

2. Hard to use: **Unsuitable formats**

- Processed data **without the raw data** that backs it, such as:
    - **Tables** on web pages,
    - **Images** / graphs
- Data embedded in **PDF**
- **Changing formats** over time (e.g. adding in new columns)

Photo by Mika Baumeister on Unsplash

2. Hard to use: **Designed for humans (only)**

- **No API** or computational access - manual human clicks required to download data, making repeatable computational pipelines difficult
- **Multiple tables in the same spreadsheet**, separated by a row or two of blank space - and **not always in a consistent order** ⭐

Photo by Mika Baumeister on Unsplash

Metadata

"The ultimate Big Data challenge lies not in the data, but in the metadata—the machine-readable descriptions that provide data about the data. It is not enough to simply put data online; data are not usable until they can be 'explained' in a manner that both humans and computers can process."

Mark Musen, Professor of Medicine, Stanford University

Quote & screenshot via metadatacenter.org

2. Hard to use: **Poorly explained / unclear data**

- Sparse, poor, or missing **metadata**
    - Poorly named and unexplained columns
    - Duplicate-named columns with no explanation
    - Raw data in hidden columns ⭐
- **Missing provenance**, possibly due to prohibitive licences
- **Data inexplicably changing** or disappearing

Source: BBC News https://www.bbc.co.uk/news/technology-54423988 (accessed August 2021)

In October 2020, nearly **16,000** coronavirus cases went **unreported** in England.

This was because the **.xls file format’s maximum column capacity** was exceeded.

2. Hard to use: **Geography changes over time**

- **District borders are regularly re-defined** at various levels of granularity (schools, elections, city boundaries, local councils, etc.)
- **Varied terms for the same place** ("UK" vs "United Kingdom")

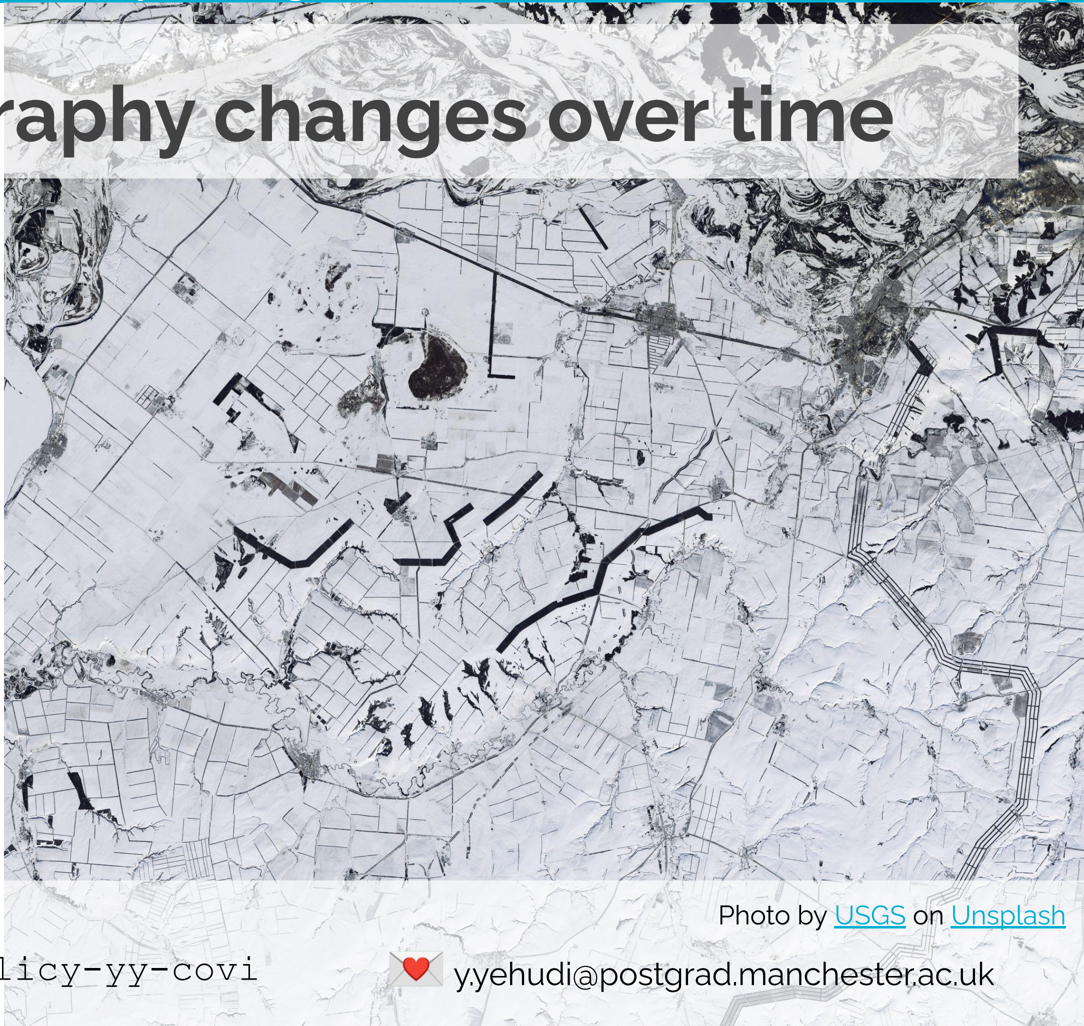


Photo by USGS on Unsplash

On approximating geography region conversions:

"Guides on how to do that from the government come with a big warning sheet saying you should **under no circumstances use this for any software that informs policy**, which is not very helpful… because we have to."

Barrier 3:

Data is **hard to re-share**

3. Hard to re-share: Types of **sharing restrictions**

---

1. Data can be **used, but not redistributed**. Leaves a big hole in the data if you redistribute your work.
2. All analyses required **approval before sharing**.
3. Often, data with **different privilege levels can't be mixed**.

On Strava, a mobility datasource:

"I spoke to one person from local government who just went "we're just not going to use it then" – **there's just no value to it if you're going to provide that amount of restriction**."

On a viral genome datasource that prohibits re-sharing:

“It’s kind of **not worth the hassle** to use that data, even though it might be informative, because if we use it we **then can’t share our data, which is a derivative**.”

Good examples & **Dream** data sources

Design for sharing **from the start**

Photo by Kyle Johnson on Unsplash

@yoyehudi

bit.ly/data4policy-yy-covid

💌 y.yehudi@postgrad.manchester.ac.uk

Good data source attributes

Design for sharing **from the start**

- **Anticipate desirable filters** - make it easy for your data users to slice by age, or region, or whatever other attributes you expect would be useful
  - This is **especially important for large datasets** - it's better to only download what you need and want!
- **Guessable URLS**, such as
  - `someurl.com/infections/location/GB-MAN.json`
  - `someurl.com/infections/age/70s.json`
- Better **metadata** and indexing.

Good data source attributes

> “You don't know which of these datasets will help with therapeutic development or not, so how can you make the judgement that this dataset is not worth helping the effort?”

- Participants spotlighted **organisations which shared data preemptively**, without needing to be asked. Examples:
  - Fine-grained **updated geographical data sets** (using new geographical boundaries)
  - Sharing **previously unreleased** viral genomes
  - Google, Apple **mobility datasets**

Good data source attributes

- Organisations that had prior **good data culture** were able to pivot quickly
  - **Data champions** who understand why good data is important
  - Data pipelines that make it easy to **swiftly add new data questions**

Wishlist items

- **Linked data** - the ability to combine multiple datasets to inform each other more effectively, e.g. linking EHR with mobility data or other data sources
- **Predictive algorithms** using big data - both for epidemic spread modelling, and for predicting viral variant behaviours.

> "Having access to highly protected things, like healthcare data, for half of my work is bizarre, because for example, we need to know what proportion of people who enter the hospital over the age of seventy dies? What is that rate? **I know it**, because I generated it for one of my models, **but I can't use it for the other model**."

Future plans

- 1-2 more interviews with targeted participants.
- Finalise analysis of results, primarily of Round 2 participants
- Create a clearer set of **recommendations for healthcare institutions, researchers, and governments**, based on the combination of barriers and good datasource/wishlist items participants have shared.

Recap

Data can be surprisingly hard to **access**, hard to **use**, and hard to **re-distribute**.

Consecutive barriers results in poorer analyses, slower predictions, mistakes, and may make existing analyses impossible to re-share.

To avoid future unnecessary deaths and long-term disability resulting from pandemics, we **must**

1. Create **technical and social data sharing pipelines** that **do not rely on human throughput**

2. Ensure that data are **well-documented with metadata**, and **well-formatted** to be readable by **humans AND by machines** (APIs).

3. **Actively share data that are not private**, whilst actively **protecting the privacy of sensitive data types**.

4. Ensure that researchers and data analysts are **free to legally re-share their analyses** and processed data, providing it does not violate privacy.

Thank you

Special thanks to:

- Interview participants,
- Everyone who helped with recruitment,
- My supervisors, Professor Caroline Jay and Professor Carole Goble

Data for Policy 2021 paper #51

Yo Yehudi

Department of Computer Science, University of Manchester

@yoyehudi

bit.ly/data4policy-yy-covid

💌 y.yehudi@postgrad.manchester.ac.uk